

Apuntes de Estadística para Ingenieros

Versión 1.3, junio de 2012

Prof. Dr. Antonio José Sáez Castillo
Dpto de Estadística e Investigación Operativa
Universidad de Jaén

Apuntes de
Estadística para Ingenieros

Prof. Dr. Antonio José Sáez Castillo
Departamento de Estadística e Investigación Operativa
Universidad de Jaén

Versión 1.3
Junio de 2012

Índice general

Prólogo

El objeto fundamental de la edición de este documento es facilitar a los alumnos de ingeniería de la Escuela Politécnica Superior de Linares el desarrollo de los contenidos teóricos de la asignatura *Estadística*. Desde un punto de vista menos local, espero que sea útil, en alguna medida, a todo aquel que necesite conocimientos básicos de las técnicas estadísticas más usuales en el ambiente científico-tecnológico.

A todos ellos, alumnos y lectores en general, quiero facilitarles el privilegio de aprender de quienes yo he aprendido, sugiriéndoles cuatro manuales que para mí han sido referencias fundamentales. Se trata, en primer lugar, del magnífico libro de Sheldon M. Ross, *Introducción a la Estadística*. En él puede encontrarse la mayor parte de lo que vamos a estudiar aquí, explicado de forma sencilla y clara, pero también comentarios históricos, reseñas bibliográficas sobre matemáticos y estadísticos relevantes y ejemplos muy apropiados. En segundo lugar, recomiendo los trabajos de William Navidi, *Estadística para ingenieros y científicos*, y Jay Devore, *Probabilidad y estadística para ingeniería y ciencias*, sobre todo por la actualidad de muchos de sus ejemplos y por cómo enfatizan el carácter aplicado, práctico, de la Estadística en el ámbito de la Ciencia y la Tecnología. Finalmente, debo mencionar también el libro de Mendenhal & Sincich, *Probabilidad y Estadística para Ingeniería y Ciencias*, que incluye, como los dos anteriores, unos ejemplos y ejercicios propuestos magníficos.

En el actual contexto del Espacio Europeo de Educación Superior, la asignatura *Estadística* tiene, en la mayor parte de los grados en ingeniería, un carácter básico y una dotación de 6 créditos ECTS. Así ocurre, por ejemplo, en las ramas de industriales o telecomunicaciones que se imparten en la Universidad de Jaén. Otras ramas, como la de ingeniería civil/minera, han optado por incluirla como asignatura obligatoria, compartida con una asignatura de ampliación de matemáticas en la que se proponen 3 créditos ECTS de estadística. Con todo, creo que estos apuntes pueden adaptarse a esos distintos contextos, aclarando qué temas pueden ser más adecuados para cada titulación. En concreto:

1. Para las distintas especialidades de la rama de industriales serían oportunos los capítulos 1, 2, 3, 4, 6, 7, 8, 9 y 10. El capítulo 9, sobre contrastes no paramétricos puede darse a modo de seminario, si el desarrollo de la docencia así lo sugiere. Sin embargo, el capítulo 10, sobre regresión lineal simple, me parece imprescindible en la formación de un futuro ingeniero industrial.

2. En los grados de la rama de telecomunicaciones, creo que son necesarios los capítulos 1, 2, 3, 4, 5, 6, 7, 8 y 11. Resulta así el temario quizá más exigente, debido a la necesidad de introducir un capítulo sobre vectores aleatorios previo a otro sobre procesos estocásticos. Queda a iniciativa del docente la posibilidad de recortar algunos aspectos en los temas tratados en aras a hacer más ligera la carga docente.

3. Finalmente, en los grados de la rama civil y minera, donde la dotación de créditos es menor, creo que

son adecuados los capítulos 1, 2, 3, 4, 6, 7, 8 y 10, si bien eliminando algunos de sus apartados, cuestión ésta que dejo, de nuevo, a juicio del docente. También sugiero que se trabajen los problemas sobre estos capítulos directamente en el contexto de unas prácticas con ordenador.

Sólo me queda pedir disculpas de antemano por las erratas que, probablemente, contienen estas páginas. Os ruego que me las hagáis llegar para corregirlas en posteriores ediciones.

Linares, junio de 2012.

Capítulo 1

Introducción

> Llegará un día en el que el razonamiento estadístico será tan necesario para el ciudadano como ahora lo es la habilidad de leer y escribir
>
> H.G. Wells (1866-1946)

Resumen. El capítulo incluye una introducción del término *Estadística* y presenta los conceptos más básicos relativos a poblaciones y muestras.

Palabras clave: estadística, población, población tangible, población conceptual, variable, muestra, muestra aleatoria simple.

1.1. ¿Qué significa Estadística?

Si buscamos en el Diccionario de la Real Academia Española de la Lengua (DRAE) el vocablo *Estadística* aparecen tres acepciones de dicha palabra[1]:

1. *Estudio de los datos cuantitativos de la población, de los recursos naturales e industriales, del tráfico o de cualquier otra manifestación de las sociedades humanas.*

2. *Conjunto de estos datos.*

3. *Rama de la matemática que utiliza grandes conjuntos de datos numéricos para obtener inferencias basadas en el cálculo de probabilidades.*

Probablemente el más común de los significados conocidos de la palabra sea el segundo, y por ello solemos ver en los medios de comunicación que cualquier recopilación de cifras referentes a algún asunto es llamado (de forma muy reduccionista) *estadística* o *estadísticas*.

Sin embargo, el valor real de la *Estadística* como ciencia tiene que ver mucho más con la primera y la tercera acepción del DRAE. Concretamente, el primero de los significados se corresponde con lo que vamos a estudiar como *Estadística Descriptiva*, donde la Estadística se utiliza para resumir, describir y explorar datos, y el tercero con lo que denominaremos *Inferencia Estadística*, donde lo que se pretende mediante la Estadística

[1] http://buscon.rae.es/draeI/SrvltGUIBusUsual?LEMA=estad %C3 %ADstica

es utilizar datos de un conjunto reducido de casos para inferir características de éstos al conjunto de todos ellos.

1.2. La Estadística en el ámbito de la Ciencia y la Ingeniería

El papel de la Estadística en la Ciencia y la Ingeniería hoy en día es crucial, fundamentalmente porque al analizar datos recopilados en experimentos de cualquier tipo, se observa en la mayoría de las ocasiones que dichos datos están sujetos a algún tipo de incertidumbre. El investigador o el profesional debe tomar decisiones respecto de su objeto de análisis basándose en esos datos, para lo cual debe dotarse de herramientas adecuadas.

A continuación vamos a describir una serie de problemas prácticos en los que se plantean situaciones de este tipo. Vamos a ponerle un nombre específico porque iremos mencionándolos a lo largo del curso, conforme seamos capaces de responder a las cuestiones que cada uno de ellos dejan abiertas.

1.2.1. Ejemplo de las capas de óxido de silicio

El artículo "Virgin Versus Recycled Wafers for Furnace Qualification: Is the Expense Justified?" (V. Czitrom y J. Reece, en *Statistical Case Studies for Industrial Process Improvement*, ASA y SIAM, 1997:87-104) describe un proceso para el crecimiento de una capa delgada de dióxido de silicio sobre placas de silicio que se usan en la fabricación de semiconductores. En él aparecen datos relativos a las mediciones del espesor, en angstroms ($\overset{\circ}{A}$), de la capa de óxido para pruebas realizadas en 24 placas: en concreto, se realizaron 9 mediciones en cada una de las 24 placas. Las placas se fabricaron en dos series distintas, 12 placas en cada serie. Estas placas eran de distintos tipos y se procesaron en distintas posiciones en el horno, ya que entre otros aspectos, el propósito de la recopilación de los datos era determinar si el espesor de la capa de óxido estaba afectado por el tipo de placa y por la posición en el horno. Por el contrario, el experimento se diseñó de tal manera que no se esperaba ninguna diferencia sistemática entre las dos series. Los datos se muestran en la Tabla 1.1.

Lo primero que salta a la vista al mirar esos datos es que es muy complicado hacerse una idea global de los resultados. Parecen estar en torno a 90 $\overset{\circ}{A}$, pero con variaciones importantes respecto de ese valor. Algunas de esas variaciones son especialmente llamativas (77.5, 106.7, ...): ¿qué pasó en esas placas? En suma, es evidente que se hace necesaria una manera sistemática de analizar los datos, tratando de describirlos de forma precisa y objetiva, respondiendo a las preguntas que subyacen en el diseño del experimento: ¿son las dos series de experimentos homogéneas? ¿afecta el tipo de placa? ¿afecta la posición en el horno? ...

1.2.2. Ejemplo de la bombilla de bajo consumo

En el envoltorio de la bombilla marca ANTE de 14W se afirma literalmente *"Lámpara ahorradora de energía. Duración 8 años"*.

Debo reconocer de que tengo mis dudas. Para empezar, ¿es que a los 8 años, de repente, la lámpara se rompe? Por otra parte, creo que todos nosotros hemos experimentado el hecho de que éstas lámparas que supuestamente tienen una duración mayor que las tradicionales lámparas incandescentes (según el envoltorio, 8 veces mayor), sin embargo, se rompen con facilidad. Luego, ¿qué quiere decir exactamente el envoltorio al afirmar que su duración es de 8 años?

Serie	Placa	Å								
1	1	90.00	92.20	94.90	92.70	91.6	88.20	92.00	98.20	96.00
1	2	91.80	94.50	93.90	77.30	92.0	89.90	87.90	92.80	93.30
1	3	90.30	91.10	93.30	93.50	87.2	88.10	90.10	91.90	94.50
1	4	92.60	90.30	92.80	91.60	92.7	91.70	89.30	95.50	93.60
1	5	91.10	89.80	91.50	91.50	90.6	93.10	88.90	92.50	92.40
1	6	76.10	90.20	96.80	84.60	93.3	95.70	90.90	100.30	95.20
1	7	92.40	91.70	91.60	91.10	88.0	92.40	88.70	92.90	92.60
1	8	91.30	90.10	95.40	89.60	90.7	95.80	91.70	97.90	95.70
1	9	96.70	93.70	93.90	87.90	90.4	92.00	90.50	95.20	94.30
1	10	92.00	94.60	93.70	94.00	89.3	90.10	91.30	92.70	94.50
1	11	94.10	91.50	95.30	92.80	93.4	92.20	89.40	94.50	95.40
1	12	91.70	97.40	95.10	96.70	77.5	91.40	90.50	95.20	93.10
2	1	93.00	89.90	93.60	89.00	93.6	90.90	89.80	92.40	93.00
2	2	91.40	90.60	92.20	91.90	92.4	87.60	88.90	90.90	92.80
2	3	91.90	91.80	92.80	96.40	93.8	86.50	92.70	90.90	92.80
2	4	90.60	91.30	94.90	88.30	87.9	92.20	90.70	91.30	93.60
2	5	93.10	91.80	94.60	88.90	90.0	97.90	92.10	91.60	98.40
2	6	90.80	91.50	91.50	91.50	94.0	91.00	92.10	91.80	94.00
2	7	88.00	91.80	90.50	90.40	90.3	91.50	89.40	93.20	93.90
2	8	88.30	96.00	92.80	93.70	89.6	89.60	90.20	95.30	93.00
2	9	94.20	92.20	95.80	92.50	91.0	91.40	92.80	93.60	91.00
2	10	101.50	103.10	103.20	103.50	˜96.1	102.50	102.00	106.70	105.40
2	11	92.80	90.80	92.20	91.70	89.0	88.50	87.50	93.80	91.40
2	12	92.10	93.40	94.00	94.70	90.8	92.10	91.20	92.30	91.10

Cuadro 1.1: Datos del espesor de las capas de óxido de silicio

En realidad, nosotros deberemos aprender a analizar este problema, asumiendo que la duración de esta bombilla no es un valor fijo y conocido, sino que está sujeto a incertidumbre. Lo que haremos será dotarnos de un modelo matemático que nos permita valorar si es probable o no que una lámpara ANTE se rompa antes de un año, después de tres años, etc.

1.2.3. Ejemplo de los niveles de plomo

Un artículo publicado en *Journal of Environmental Engineering* en 2002, titulado "Leachate from Land Disposed Residential Construction Waste", presenta un estudio de la contaminación en basureros que contienen desechos de construcción y desperdicios de demoliciones. De un sitio de prueba se tomaron 42 muestras de lixiado, de las cuales 26 contienen niveles detectables de plomo. Se pone así de manifiesto que sólo una parte de los basureros está contaminada por plomo. La cuestión es ¿qué proporción supone esta parte contaminada de la superficie total de los basureros?

Si una ingeniera desea obtener a partir de esos datos una estimación de la proporción de los basureros que contiene niveles detectables de plomo debe ser consciente de dos cuestiones:

1. Es imposible analizar todos los rincones de todos los basureros.

2. Si se basa sólo en los datos del artículo, esa estimación será sólo eso, una estimación basada en esa muestra, que es de sólo 42 datos. Debería, por tanto obtener también una estimación del error que está cometiendo al hacer la estimación. Con ambos resultados, la estimación en sí y una cuantificación del error que podría cometer con ella, incluso podrá obtener un rango donde la verdadera proporción se encuentra, con un alto nivel de confianza.

1.2.4. Ejemplo de los cojinetes

Un ingeniero industrial es responsable de la producción de cojinetes de bolas y tiene dos máquinas distintas para ello. Le interesa que los cojinetes producidos tengan diámetros similares, independientemente de la máquina que los produce, pero tiene sospechas de que está produciendo algún problema de falta de calibración entre ellas. Para analizar esta cuestión, extrae una muestra de 120 cojinetes que se fabricaron en la máquina A, y encuentra que la media del diámetro es de 5.068 mm y que su desviación estándar es de 0.011 mm. Realiza el mismo experimento con la máquina B sobre 65 cojinetes y encuentra que la media y la desviación estándar son, respectivamente, 5.072 mm y 0.007 mm. ¿Puede el ingeniero concluir que los cojinetes producidos por las máquinas tienen diámetros medios significativamente diferentes?

1.2.5. Ejemplo de la absorción de un compuesto a distintas dosis y en distintos tiempos de absorción

Un equipo de investigadores que trabajan en seguridad en el trabajo está tratando de analizar cómo la piel absorbe un cierto componente químico peligroso. Para ello, coloca diferentes volúmenes del compuesto químico sobre diferentes segmentos de piel durante distintos intervalos de tiempo, midiendo al cabo de ese tiempo el porcentaje de volumen absorbido del compuesto. El diseño del experimento se ha realizado para que la interacción esperable entre el tiempo y el volumen no influya sobre los resultados. Los datos se mostrarán en el último tema.

Lo que los investigadores se cuestionan es si la cantidad de compuesto por un lado y el tiempo de exposición al que se somete por otro, influyen en el porcentaje que se absorbe. De ser así, sería interesante estimar el porcentaje de absorción de personas que se sometan a una exposición de una determinada cantidad, por ejemplo, durante 8 horas.

1.2.6. Ejemplo de los accidentes laborales

En una empresa se sospecha que hay franjas horarias donde los accidentes laborales son más frecuentes. Para estudiar este fenómeno, contabilizan los accidentes laborales que sufren los trabajadores según franjas horarias, durante un año. Los resultados aparecen en la tabla.

Horas del día	Número de accidentes
8-10 h.	47
10-12 h.	52
13-15 h.	57
15-17 h.	63

Con esa información, los responsables de seguridad de la empresa deben decidir si hay franjas horarias donde los accidentes son más probables o si, por el contrario, éstos ocurren absolutamente al azar.

1.2.7. Ejemplo de la cobertura de la antena de telefonía móvil

Reduciendo mucho el problema, supongamos que una antena de telefonía móvil tiene una cobertura que abarca a cualquier móvil dentro de un círculo de radio r. Un ingeniero puede suponer que un teléfono concreto puede estar situado *en cualquier punto al azar* de ese círculo, pero ¿cómo plasmar eso? Por ejemplo, si nos centramos en la distancia a la antena, ¿cualquier distancia es *igualmente probable*? ¿Y qué podemos decir de las coordenadas en un momento concreto del móvil?

1.2.8. Ejemplo de la señal aleatoria

En el contexto de las telecomunicaciones, cualquier señal debe considerarse aleatoria, es decir, debe tenerse en cuenta que cuando la observamos, parte de ella es debida a la incertidumbre inherente a cualquier proceso de comunicación. Y es que, por multitud de razones, nadie tiene garantías que la señal enviada sea exactamente igual a la señal recibida.

Un ingeniero debe tener en cuenta eso y, a pesar de todo, ser capaz de analizar las propiedades más relevantes de cualquier señal y de estudiar su comportamiento en cualquier momento del proceso de comunicación.

Por ejemplo, hoy en día una señal sufre multitud de transformaciones en el proceso de comunicación. Cada una de esas transformaciones se considera el resultado del paso de la señal por un sistema. El ingeniero debe ser capaz de conocer las características más relevantes de la señal a lo largo de todas esas transformaciones.

1.3. Definiciones básicas

Para finalizar este primer tema de introducción, vamos a ir fijando las definiciones más elementales que utilizaremos a lo largo del curso y que ya han sido motivadas en la introducción de los ejemplos anteriores.

Se denomina **población** a un conjunto de individuos o casos, objetivo de nuestro interés.

Podemos distinguir entre poblaciones tangibles y poblaciones conceptuales.

Una población es **tangible** si consta de elementos físicos reales que forman un conjunto finito.

Por ejemplo, si estamos considerando el estudio de la altura de los alumnos de la Escuela, el conjunto de estos alumnos es una población tangible.

Una población **conceptual** no tiene elementos reales, sino que sus casos se obtienen por la repetición de un experimento.

Por ejemplo, cuando planteábamos las pruebas sobre placas de silicio, vemos que hay tantos casos como pruebas puedan hacerse, lo que supone un conjunto infinito de casos. En poblaciones conceptuales es imposible, por tanto, conocer todos los casos, y tenemos que conformarnos con muestras de los mismos.

Una **variable** o **dato** es una característica concreta de una población.

Por ejemplo:

- Si consideramos la población de todos los alumnos de la Escuela, podemos fijarnos en la variable *altura*.
- Si consideramos el supuesto de las pruebas sobre placas de silicio, podemos considerar la variable *espesor de la capa de óxido de silicio generada*.

Se denomina **muestra** a cualquier subconjunto de datos seleccionados de una población.

El objetivo de una muestra, ya sea en una población tangible o en una población conceptual es que los elementos de la muestra **representen** al conjunto de todos los elementos de la población. Esta cuestión, la construcción de muestras adecuadas, representativas, es uno de los aspectos más delicados de la Estadística.

Nosotros vamos a considerar en esta asignatura sólo un tipo de muestras, denominadas **muestras aleatorias simples**. En una muestra aleatoria simple, todos los elementos de la población deben tener las mismas posibilidades de salir en la muestra y, además, los elementos de la muestra deben ser independientes: el que salga un resultado en la muestra no debe afectar a que ningún otro resultado salga en la muestra.

Por ejemplo, podríamos estar interesados en la población de todos los españoles con derecho a voto (población tangible, pero enorme), de los que querríamos conocer un dato o variable, su intención de voto en las próximas elecciones generales. Dado que estamos hablando de millones de personas, probablemente deberemos escoger una muestra, es decir, un subconjunto de españoles a los que se les realizaría una encuesta. Si queremos que esa muestra sea aleatoria simple, deberemos tener cuidado de que todos los españoles con derecho a voto tengan las mismas posibilidades de caer en la muestra y de que la respuesta de un entrevistado no afecte a la de ningún otro. Como nota curiosa, sabed que la mayoría de las encuestas nacionales se hacen vía telefónica, lo cual es una pequeña violación de las hipótesis de muestra aleatoria simple, ya que hay españoles con derecho a voto que no tienen teléfono, luego es imposible que salgan en la muestra.

Parte I

Estadística descriptiva

Capítulo 2

El tratamiento de los datos. Estadística descriptiva

> Es un error capital el teorizar antes de poseer datos. Insensiblemente uno comienza a alterar los hechos para encajarlos en las teorías, en lugar encajar las teorías en los hechos
>
> Sherlock Holmes (A. C. Doyle), en *Un escándalo en Bohemia*

Resumen. En este capítulo aprenderemos métodos para resumir y describir conjuntos de datos a través de distintos tipos de tablas, gráficos y medidas estadísticas.

Palabras clave: datos cuantitativos, datos cualitativos, datos discretos, datos continuos, distribución de frecuencias, diagrama de barras, diagrama de sectores, histograma, media, mediana, moda, cuantiles, varianza, desviación típica, asimetría, datos atípicos.

2.1. Introducción

Obtenidos a través de encuestas, experimentos o cualquier otro conjunto de medidas, los datos estadísticos suelen ser tan numerosos que resultan prácticamente inútiles si no son resumidos de forma adecuada. Para ello la Estadística utiliza tanto técnicas gráficas como numéricas, algunas de las cuales describimos en este capítulo.

Podemos decir que existe una clasificación, un tanto artificial, de los datos, según se refieran a una población tangible, en cuyo caso se conocerán todos los casos, o a una población conceptual, en cuyo caso sólo se conocerá una muestra (aleatoria simple). Sin embargo, esta clasificación no tiene ningún efecto en lo relativo a lo que vamos a estudiar en este capítulo.

2.2. Tipos de datos

Los datos (o variables) pueden ser de dos tipos: **cuantitativos** y **cualitativos.**

Los datos **cuantitativos** son los que representan una cantidad reflejada en una escala numérica. A su vez, pueden clasificarse como datos **cuantitativos discretos** si se refieren al conteo de alguna característica, o datos **cuantitativos continuos** si se refieren a una medida.

Los datos **cualitativos o categóricos** se refieren a características de la población que no pueden asociarse a cantidades con significado numérico, sino a características que sólo pueden clasificarse.

Ejemplo. Veamos algunos ejemplos de cada uno de estos tipos de variables:

- En el ejemplo del óxido de silicio, la variable *espesor* es cuantitativa continua.
- En el ejemplo de los cojinetes, el *diámetro de los cojinetes* es una variable cuantitativa continua.
- En el ejemplo de los niveles de plomo, se está analizando si una muestra contiene niveles detectables o no. Se trata, por tanto, de una variable cualitativa con dos categorías: *sí contiene niveles detectables* o *no contiene niveles detectables*.
- En el ejemplo de los accidentes laborales, la variable *número de accidentes laborales* es cuantitativa discreta, mientras que las franjas horarias constituyen una variable cualitativa.

2.3. Métodos gráficos y numéricos para describir datos cualitativos

La forma más sencilla de describir de forma numérica una variable cualitativa es determinar su distribución de frecuencias. Por su parte, esta distribución de frecuencias determina a su vez las representaciones gráficas más usuales.

Supongamos que tenemos una variable cualitativa, que toma una serie de posibles valores (categorías). El número de veces que se da cada valor es la **distribución de frecuencias** de la variable. Si en vez de dar el número de veces nos fijamos en la proporción de veces, tenemos la **distribución de frecuencias relativas**.

Las representaciones gráficas más usuales son los diagramas de barras y los diagramas de sectores.

Los **diagramas de barras** son una representación de cada una de las categorías de la variable mediante una barra colocada sobre el eje X y cuya altura sea la frecuencia o la frecuencia relativa de dichas categorías.

Los **diagramas de sectores** son círculos divididos en tantos sectores como categorías, sectores cuyo ángulo debe ser proporcional a la frecuencia de cada categoría.

Categoría	Frecuencia	Frecuencia relativa
País	Número de reactores nucleares	Proporción
Bélgica	4	0.041
Francia	22	0.225
Finlandia	2	0.020
Alemania	7	0.071
Holanda	1	0.010
Japón	11	0.112
Suecia	3	0.031
Suiza	1	0.010
Estados Unidos	47	0.480
TOTAL	98	1.000

Cuadro 2.1: Tabla de frecuencias.

Ejemplo. Tomamos como población los 98 reactores nucleares más grandes en todo el mundo. Nos fijamos en la variable o dato referente al país donde están localizados.

Los datos serían

Bélgica, Bélgica, Bélgica, Bélgica, Francia, Finlandia, Finlandia, Alemania, Alemania, Alemania, Alemania, Alemania, Alemania, Alemania, Holanda, Japón, Japón, Japón, Japón, Japón, Japón, Japón, Japón, Japón, Japón, Japón, Suecia, Suecia, Suecia, Suiza, Estados Unidos.

Las distribuciones de frecuencias y de frecuencias relativas podemos resumirlas en una **tabla de frecuencias** como la que aparece en el Cuadro 2.1.

Por su parte, las representaciones mediante diagramas de barras y sectores de estos datos aparecen en la Figura 2.1 y la Figura 2.2 respectivamente.

2.4. Métodos gráficos para describir datos cuantitativos

Si tenemos una variable cuantitativa discreta y ésta toma pocos valores, podemos tratarla como si fuera una variable cualitativa, calcular su distribución de frecuencias y dibujar un diagrama de barras.

Ejemplo. En una empresa con cadena de montaje donde se empaquetan piezas en cajas se realiza un estudio sobre la calidad de producción. Los datos siguientes informan sobre el número de piezas defectuosas encontradas en una muestra de cajas examinadas:

0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 5 5 5 5 6 6 6 6 6 7 7 7 8 8 9

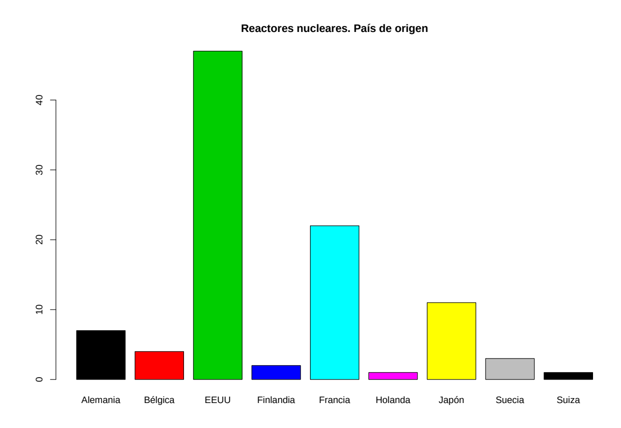


Figura 2.1: Diagrama de barras.

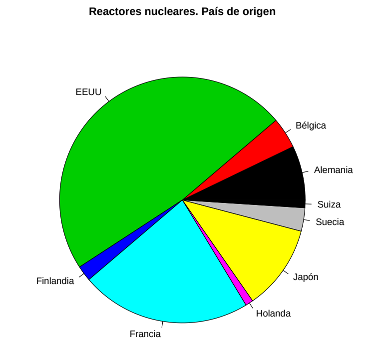


Figura 2.2: Diagrama de sectores.

El diagrama de barras asociado aparecen en la Figura 2.3.

Sin embargo, la mayoría de variables cuantitativas son de tipo continuo, de manera que toman demasiados valores como para que la representación de su distribución de frecuencias sea útil[1]. Por ello el método gráfico más común y tradicional para datos cuantitativos es el histograma.

El **histograma** es una variante del diagrama de barras donde se agrupan los valores de la variable en intervalos para que estos intervalos tengan frecuencias mayores que uno.

Para obtener un histograma de forma manual deben seguirse los siguientes pasos:

1. Calculamos el número, N, de intervalos que vamos a utilizar. Se recomienda que sea aproximadamente igual a la raíz cuadrada del número de datos. Sin embargo, los programas estadísticos suelen utilizar otro método, llamado *Método de Sturges*, en el que $N = \lceil \log_2 n + 1 \rceil$, donde n es el número de datos y $[]$ es la función parte entera.

[1] Si toma muchos valores, muy probablemente la mayor parte de ellos sólo aparezca una vez, por lo que la distribución de frecuencias será casi siempre constante e igual a 1.

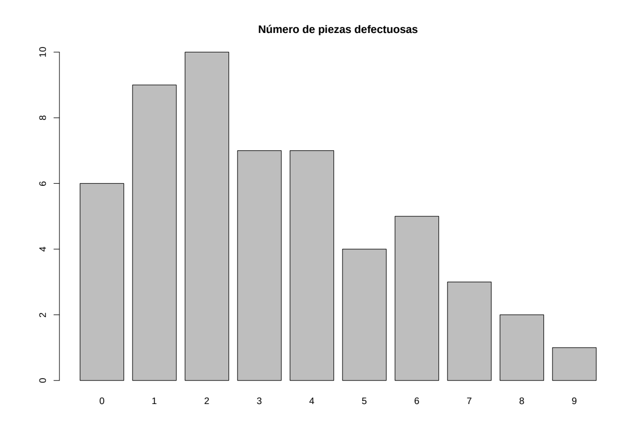


Figura 2.3: Diagrama de barras.

2. Calculamos el rango, R, del histograma, que será ligeramente más amplio que el rango de los datos. El histograma debe comenzar en un número (x_m) ligeramente por debajo del mínimo de los datos y terminar en un número (x_M) ligeramente por encima del máximo. El rango del histograma será, por tanto, $R = x_M - x_m$.

3. Calculamos la longitud, L, de los intervalos, como el cociente entre el rango del histograma y el número de intervalos, es decir, $L = \frac{R}{N}$.

4. Se construyen los N intervalos:

$$I_1 = [x_m, x_m + L)$$
$$I_2 = [x_m + L, x_m + 2L)$$
$$I_3 = [x_m + 2L, x_m + 3L)$$
$$\ldots$$
$$I_N = [x_m + N \times L, x_M).$$

5. Para cada intervalo, contamos el número de datos que hay en él, es decir, la frecuencia del intervalo.

6. El histograma es un diagrama de barras donde en el eje X se colocan los intervalos y sobre ellos se construyen barras cuya altura sea la frecuencia o la frecuencia relativa del intervalo. En este caso, las barras deben dibujarse sin espacio entre ellas. En ocasiones, en vez de tomar la frecuencia relativa como altura de las barras, se toma dicha frecuencia relativa como área de las barras: en ese caso, se habla de un histograma en escala de densidad.

Nota. Por cuestiones que detallaremos más adelante es importante destacar que el porcentaje de datos que cae dentro de un intervalo es proporcional al área de la barra que se construye sobre ese intervalo. Por ejemplo, si el área de una barra es el 30 % del área total del intervalo, entonces el 30 % de los datos están en dicho intervalo.

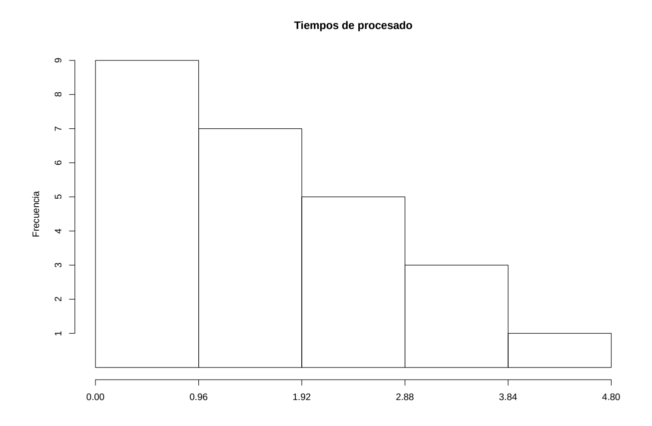


Figura 2.4: Histograma.

Por otra parte, ¿qué pasaría si tomamos un número muy grande de datos? El número de intervalos del histograma sería también muy grande, y las barras serían muy estrechas, de manera que en vez de parecer un diagrama de barras, parecería la gráfica de una función real de variable real. Hablaremos de esta función y del área debajo de ella en breve. Por cierto, ¿cómo se calcula el área bajo esta función?

Ejemplo. Los datos siguientes corresponden al tiempo necesario para procesar 25 trabajos en una CPU.

1.17	1.61	1.16	1.38	3.53	1.23	3.76	1.94	0.96	4.75
0.15	2.41	0.71	0.02	1.59	0.19	0.82	0.47	2.16	2.01
0.92	0.75	2.59	3.07	1.4					

Vamos a calcular un histograma para esos datos.

1. Dado que $\sqrt{25} = 5$, utilizaremos 5 intervalos.

2. El mínimo de los datos es 0.02 y el máximo 4.75, de manera que podemos considerar como rango del histograma el intervalo $[0, 4.8]$, cuya longitud (rango del histograma) es 4.8.

3. La longitud de los intervalos es, en ese caso, $\frac{4.8}{5} = 0.96$.

4. Construimos los intervalos:

$$I_1 = [0, 0.96)$$
$$I_2 = [0.96, 1.92)$$
$$I_3 = [1.92, 2.88)$$
$$I_4 = [2.88, 3.84)$$
$$I_5 = [3.84, 4.8)$$

5. Calculamos la distribución de frecuencia asociada a esos intervalos:

Tiempo de procesado	Frecuencia
$[0, 0.96)$	8
$[0.96, 1.92)$	8
$[1.92, 2.88)$	5
$[2.88, 3.84)$	3
$[3.84, 4.8)$	1

6. Finalmente, representamos el diagrama de barras (Figura 2.4).

2.5. Métodos numéricos para describir datos cuantitativos

Es cierto que un diagrama de barras o un histograma nos ayudan a tener una imagen de cómo son los datos, pero normalmente es necesario complementar esa imagen mediante medidas que, de forma objetiva, describan las características generales del conjunto de datos.

Vamos a ver en este apartado tres tipos de medidas, que básicamente responden a tres preguntas: *por dónde están los datos* (medidas de posición), *cómo de agrupados están los datos* (medidas de dispersión) y *qué forma tienen los datos* (medidas de forma).

2.5.1. Medidas de tendencia central

Las **medidas de tendencia central** son medidas de posición que tratan de establecer un valor que pueda considerarse *el centro* de los datos en algún sentido.

2.5.1.1. Media

Sea un conjunto de datos de una variable cuantitativa, $x_1, ..., x_n$. La **media** de los datos es

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Esta medida es la más común dentro de las de tendencia central y corresponde al *centro de gravedad* de los datos.

Es inmediato comprobar que si se realiza un cambio de origen y escala sobre los datos, del tipo $y = ax + b$, la media sufre el mismo cambio, es decir, $\bar{y} = a\bar{x} + b$.

De igual forma, si tenemos datos de la suma de dos o más variables, la media de la suma es la suma de las medias de cada variable.

2.5.1.2. Mediana

Sea un conjunto de datos de una variable cuantitativa, $x_1, ..., x_n$. Ordenemos la muestra de menor a mayor, $x_{(1)}, ..., x_{(n)}$.

La **mediana** es el valor de la variable que deja el mismo número de datos antes y después que él, una vez ordenados estos.

El cálculo de la mediana dependerá de si el número de datos, n, es par o impar:

- Si n es impar, la mediana es el valor que ocupa la posición $\frac{n+1}{2}$ una vez que los datos han sido ordenados (en orden creciente o decreciente), porque éste es el valor central. Es decir: $M_e = x_{\left(\frac{n+1}{2}\right)}$.
- Si n es par, la mediana es la media aritmética de las dos observaciones centrales. Cuando n es par, los dos datos que están en el centro de la muestra ocupan las posiciones $\frac{n}{2}$ y $\frac{n}{2}+1$. Es decir: $M_e = \frac{x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}}{2}$.

La mediana corresponde exactamente con la idea de valor central de los datos. De hecho, puede ser un valor más representativo de éstos que la media, ya que es más *robusta* que la media. Veámos qué significa esto en un ejemplo.

Ejemplo. Consideremos los datos siguientes:

0 0 1 2 3 4 5

Su media es $\frac{0+0+1+2+3+4+5}{7} = 2.1429$, y su mediana 2.

Pero imaginemos que por error o por casualidad obtenemos un nuevo dato enormemente grande en relación al resto de datos, 80. En ese caso, la media sería

$$\frac{0+0+1+2+3+4+5+80}{8} = 11.875$$

y la mediana 2.5. Es decir, un solo dato puede desplazar enormemente la media, hasta convertirla en una medida poco representativa, pero sólo desplazará ligeramente la mediana. Ese es el motivo por el que se dice que la mediana es una medida **robusta**.

2.5.1.3. Moda o intervalo modal

En principio la **moda** se define como el valor más frecuente de los datos. Lo que ocurre es que si éstos son datos de una variable continua o discreta con muchos valores, puede que los datos apenas se repitan. En ese caso, en el que, como vimos en las representaciones gráficas, se debe agrupar por intervalos, no debe darse un valor como moda, sino un **intervalo modal**, aquél con mayor frecuencia asociada.

2.5.2. Cuantiles

Los **cuantiles** son medidas de posición pero no necesariamente ligados al *centro* de los datos. La idea a la que responden es muy sencilla y muy práctica. Se trata de valorar de forma relativa cómo es un dato respecto del conjunto global de todos los datos.

Si, por ejemplo, un niño de 4 años pesa 13 kilos, ¿está desnutrido? ¿está sano? La respuesta debe ser que *depende*. ¿Dónde vive el niño? Es importante porque, por ejemplo, en Estados Unidos los niños son en general más grandes que, por ejemplo, en Japón. Quizá más que el peso nos interese saber qué posición relativa tiene el peso del niño dentro de la población de la que forma parte. Por ejemplo, si nos dicen que el niño está entre el 1 % de los niños que menos pesan, probablemente tiene un problema de crecimiento.

El **cuantil** p (Q_p) de unos datos ($0 \leq p \leq 1$), sería un valor de la variable situado de modo que el $100p\,\%$ de los valores sean menores o iguales que él y el resto ($100(1-p)\,\%$) mayores.

No obstante, en la práctica vamos a encontrar un problema para encontrar cuantiles, sobre todo con pocos datos: lo más habitual es que no exista el valor exacto que deje a la izquierda el $100p\,\%$ de los valores y el resto a la derecha. Por ese motivo, los programas estadísticos utilizan unas fórmulas de interpolación para obtener el valor del cuantil entre los dos valores de los datos que lo contienen. En nuestro caso, a la hora de obtener cuantiles, la aplicación de esas fórmulas de interpolación *a mano* harían muy lentos y pesados los cálculos, por lo que vamos a aplicar un convenio mucho más sencillo: aproximaremos el valor del cuantil correspondiente de la siguiente forma:

1. Si el $100p\,\%$ de n, donde n es el número de datos, es un entero, k, entonces $Q_p = \frac{x_{(k)}+x_{(k+1)}}{2}$.
2. Si el $100p\,\%$ de n no es un entero, lo redondeamos al entero siguiente, k, y entonces $Q_p = x_{(k)}$.

No olvidemos, sin embargo, que los programas estadísticos van a utilizar las fórmulas de interpolación para calcular el valor de los cuantiles, de manera que no debe extrañar si se observan pequeñas diferencias al comparar nuestros resultados *a mano* con los de estos programas.

Existen diversos nombres para referirse a algunos tipos de cuantiles. Entre ellos:

- Los **percentiles** son los cuantiles que dividen la muestra en 100 partes, es decir, son los cuantiles 0.01 (percentil 1), 0.02 (percentil 2), ..., 0.99 (percentil 99). Si notamos por P_α al percentil α, con $\alpha = 1, 2, 3, ..., 99$, se tiene que $P_\alpha = Q_{\alpha/100}$. En Estadística Descriptiva es más frecuente hablar de percentiles que de cuantiles porque se refieren a cantidades entre 0 y 100, en tanto por ciento, que son más habituales de valorar por todo el mundo.
- Los **cuartiles** dividen a la población en cuatro partes iguales, es decir, corresponden a los cuantiles 0.25, 0.5 (mediana) y 0.75.

Ejemplo. Consideremos de nuevo los datos correspondientes al tiempo de procesado de 25 tareas en una CPU. Ahora los hemos ordenado de menor a mayor (en 5 filas):

0.02	0.75	1.17	1.61	2.59
0.15	0.82	1.23	1.94	3.07
0.19	0.92	1.38	2.01	3.53
0.47	0.96	1.40	2.16	3.76
0.71	1.16	1.59	2.41	4.75

Vamos a calcular distintas medidas de posición y a comentarlas.

En primer lugar, la media es 1.63. La mediana ocupa el lugar 13 en la muestra ordenada, y su valor es 1.38. Obsérvese que la media es algo mayor que la mediana: esto es debido a la presencia de algunos valores significativamente más altos que el resto, como pudimos ver en el histograma.

Por su parte, el P_{25} o cuantil 0.25 ocupa la posición 7, ya que el 25 % de 25 es 6.25. Por tanto, $P_{25} = 0.82$. De igual forma, $P_{75} = Q_{0.75} = 2.16$, el valor que ocupa la posición 19. Podemos ver, por tanto, que los valores más bajos están muy agrupados al principio, y se van dispersando más conforme se hacen más altos.

2.5.3. Medidas de variación o dispersión

Las **medidas de variación o dispersión** están relacionadas con las medidas de tendencia central, ya que lo que pretenden es cuantificar cómo de concentrados o dispersos están los datos respecto a estas medidas. Nosotros nos vamos a limitar a dar medidas de dispersión asociadas a la media.

La idea de estas medidas es valorar en qué medida los datos están agrupados en torno a la media. Esta cuestión tan simple es uno de los motivos más absurdos de la mala prensa que tiene la Estadística en la sociedad en general. La gente no se fía de lo que ellos llaman *la Estadística* entre otros motivos, porque parece que todo el mundo cree que una media tiene que ser un valor válido para todos, y eso es materialmente imposible.

Ejemplo. Pensemos en la media del salario de los españoles. En 2005 fue de 18.750 euros al año. Ahora bien, esa media incluye tanto a las regiones más desarrolladas como a las más desfavorecidas y, evidentemente, la cifra generará mucho malestar en gran parte de la población (con toda seguridad, más del 50 %), cuyo salario está por debajo.

Ejemplo. Existe una frase muy conocida que dice que "*la Estadística es el arte por el cuál si un español se come un pollo y otro no se come ninguno, se ha comido medio pollo cada uno*". Esa frase se usa en muchas ocasiones para ridiculizar a la Estadística, cuando en realidad debería servir para desacreditar a quien la dice, por su ignorancia.

Hay que decir que la Estadística no tiene la culpa de que la gente espere de una media más de lo que es capaz de dar, ni de que muy poca gente conozca medidas de dispersión asociadas a la media.

2.5.3.1. Varianza muestral

Dados unos datos de una variable cuantitativa, $x_1, ..., x_n$, la **varianza muestral**[2] de esos datos es

$$s_{n-1}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}.$$

Nota. Para calcular *a mano* la varianza resulta más cómodo desarrollar un poco su fórmula, como vamos a ver:

$$s_{n-1}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - 2\bar{x}\sum_{i=1}^n x_i + n\bar{x}^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2}{n-1}$$
$$= \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1}.$$

Cuanto mayor sea la varianza de unos datos, más dispersos, heterogéneos o variables son esos datos. Cuanto más pequeña sea una varianza de unos datos, más agrupados u homogéneos son dichos datos.

Ejemplo. Una muestra aleatoria simple de la altura de 5 personas arroja los siguientes resultados:

$$1.76 \quad 1.72 \quad 1.80 \quad 1.73 \quad 1.79$$

Calculemos su media y su varianza muestral.

Lo único que necesitamos es $\sum_{i=1}^5 x_i = 8.8$ y $\sum_{i=1}^5 x_i^2 = 15.493$. A partir de estos datos,

$$\bar{x} = \frac{8.8}{5} = 1.76$$

y

$$s_{n-1}^2 = \frac{15.493 - 5 \times 1.76^2}{4} = 0.00125$$

En lo que respecta al comportamiento de la varianza muestral frente a cambios de origen y escala, sólo le afectan los segundos. Es decir, si tenemos que $y = ax + b$, se verifica que $s_{y;n-1}^2 = a^2 s_{x;n-1}^2$.

Finalmente, si bien habíamos comentado que en el caso de la media, si tenemos la suma de varias variables, la media total es la suma de las medias de cada variable, no ocurre así con la varianza en general.

2.5.3.2. Desviación típica o estandar muestral

El principal problema de la varianza es su unidad de medida. Por cómo se define si, por ejemplo, la variable se expresa en kilos, la media también se expresa en kilos, pero la varianza se expresa en kilos^2, lo que hace que sea difícil valorar si una varianza es muy elevada o muy pequeña.

Es por ello que se define la **desviación típica o estandar muestral** de los datos como $s_{n-1} = \sqrt{s_{n-1}^2}$, cuya unidad de medida es la misma que la de la media.

Nota. La Regla Empírica

Si el histograma asociado a unos datos tiene la forma de una campana o de una joroba, el conjunto de datos tendrá las siguientes características, lo que en algunos libros se conoce como **Regla Empírica**:

1. Aproximadamente el 68 % de los datos estará en el intervalo $(\bar{x} - s_{n-1}, \bar{x} + s_{n-1})$.
2. Aproximadamente el 95 % de los datos estará en el intervalo $(\bar{x} - 2s_{n-1}, \bar{x} + 2s_{n-1})$.
3. Casi todos los datos estarán en el intervalo $(\bar{x} - 3s_{n-1}, \bar{x} + 3s_{n-1})$.

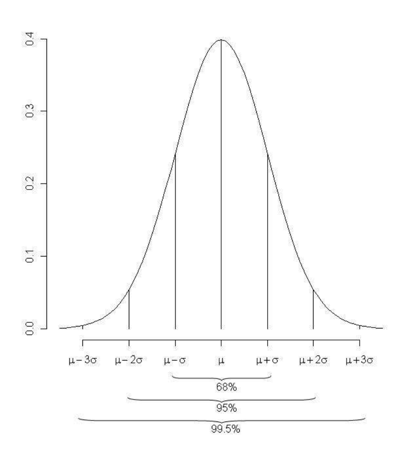


Figura 2.5: Representación gráfica de la regla empírica.

2.5.3.3. Coeficiente de variación

Como acabamos de decir, debemos proporcionar cada media junto con alguna medida de dispersión, preferentemente la desviación típica. Una forma de valorar en términos relativos cómo es de dispersa una variable es precisamente proporcionar el cociente entre la desviación típica y la media (en valor absoluto), lo que se conoce como **coeficiente de variación.**

Dado un conjunto de datos de media $\bar{x}$ y desviación típica s_{n-1}, se define su **coeficiente de variación** como

$$CV = \frac{s_{n-1}}{|\bar{x}|}.$$

La principal ventaja del coeficiente de variación es que no tiene unidades de medida, lo que hace más fácil su interpretación.

Ejemplo. Para los datos de tiempo de procesado en una CPU de 25 tareas, la varianza es 1.42, luego su desviación estandar es 1.19, y el coeficiente de variación $\frac{1.19}{1.63} = 0.73$. Por tanto, la desviación estándar es algo más del 70 % de la media. Esto indica que los datos no están muy concentrados en torno a la media, probablemente debido a la presencia de los valores altos que hemos comentado antes.

Nota. El coeficiente de variación, tal y como está definido, sólo tiene sentido para conjuntos de datos con el mismo signo, es decir, todos positivos o todos negativos. Si hubiera datos de distinto signo, la media podría estar próxima a cero o ser cero, imposibilitando que aparezca en el denominador.

Nota. Suele ser frecuente el error de pensar que el coeficiente de variación no puede ser mayor que 1, lo cual es rigurosamente falso. Si lo expresamos en porcentaje, el coeficiente de variación puede ser superior al 100 % sin más que la desviación típica sea mayor que la media, cosa bastante frecuente, por cierto.

Nota. A la hora de interpretar el coeficiente de variación inmediatamente surge la pregunta de *¿cuándo podemos decir que es alto y cuándo que es bajo?* Realmente, no existe una respuesta precisa, sino que depende del contexto de los datos que estemos analizando. Si, por ejemplo, estamos analizando unos datos que por su naturaleza deben ser muy homogéneos, un coeficiente de variación del 10 % sería enorme, pero si por el contrario estamos analizando datos que por su naturaleza son muy variables, un coeficiente de variación del 10 % sería muy pequeño.

Por todo ello, lo recomendable es analizar el coeficiente de variación entendiendo su significado numérico, es decir, entendiendo que se refiere a la comparación de la desviación típica con la media, e interpretando su valor en relación al contexto en el que estemos trabajando.

2.5.4. Medidas de forma. Coeficiente de asimetría

Las **medidas de forma** comparan la forma que tiene la representación gráfica, bien sea el histograma o el diagrama de barras de la distribución, con una situación *ideal* en la que los datos se reparten en igual medida a la derecha y a la izquierda de la media.

Esa situación en la que los datos están repartidos de igual forma a uno y otro lado de la media se conoce como **simetría**, y se dice en ese caso que la distribución de los datos es simétrica. En ese caso, además, su mediana, su moda y su media coinciden.

Por contra, se dice que una distribución es **asimétrica a la derecha** si las frecuencias (absolutas o relativas) descienden más lentamente por la derecha que por la izquierda. Si las frecuencias descienden más lentamente por la izquierda que por la derecha diremos que la distribución es **asimétrica a la izquierda**.

Para valorar la simetría de unos datos se suele utilizar el **coeficiente de asimetría de Fisher**:

$$As = \frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{n-1}}{s_{n-1}^3}.$$

Obsérvese que para evitar el problema de la unidad y hacer que la medida sea escalar y por lo tanto relativa, dividimos por el cubo de su desviación típica. De esta forma podemos valorar si unos datos son más o menos simétricos que otros, aunque no estén medidos en la misma unidad de medida. La interpretación de este coeficiente de asimetría es la siguiente:

- Tanto mayor sea el coeficiente en valor absoluto, más asimétricos serán los datos.
- El signo del coeficiente nos indica el sentido de la asimetría:
  - Si es positivo indica que la asimetría es a la derecha.
  - Si es negativo, indica que la asimetría es a la izquierda.

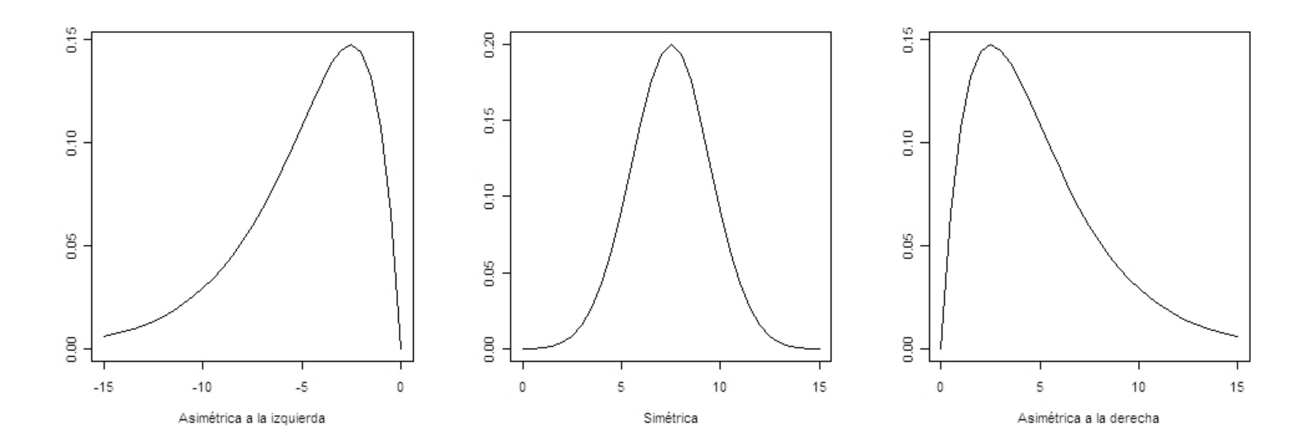


Figura 2.6: Formas típicas de distribuciones de datos.

Ejemplo. Para los datos de tiempo de procesado en una CPU de 25 tareas, el coeficiente de asimetría de Fisher es 0.91, lo que, como habíamos visto y comentado con anterioridad, pone de manifiesto que la distribución es asimétrica a la derecha, debido a la presencia de tiempos de procesado bastante altos en relación al resto.

2.5.5. Parámetros muestrales y parámetros poblacionales

Cuando se trabaja con una muestra de una población, ya sea ésta tangible o conceptual, las distintas medidas de posición, dispersión y forma, se denominan **parámetros muestrales**. Hay que tener en cuenta que prácticamente siempre se trabaja con muestras, ya que o bien trabajamos con poblaciones conceptuales o con poblaciones tangibles (finitas, por tanto), pero con muchísimos elementos.

Frente a estos parámetros muestrales se encuentran los parámetros análogos referidos a toda la población. Estos parámetros, llamados **parámetros poblacionales**, son, en general, imposibles de conocer[3]. Por ejemplo, la media poblacional se calcularía igual que la media muestral de unos datos, pero aplicada la fórmula a todos los elementos de la población. Como eso es prácticamente imposible de poner en la práctica, veremos

[3] Salvo en el caso de poblaciones finitas con pocos elementos.

en capítulos posteriores que los parámetros muestrales se utilizan en la práctica para aproximar o estimar los parámetros poblacionales.

2.6. Métodos para detectar datos cuantitativos atípicos o fuera de rango

Hay ocasiones en que un conjunto de datos contiene una o más observaciones *inconsistentes* en algún sentido. Por ejemplo, en los datos de tiempo de procesado en una CPU de 25 tareas, supongamos que tenemos una observación más, igual a 85, debido a que la CPU se bloqueó y hubo que reiniciarla. Este dato, que probablemente no deseemos incluir, es un ejemplo de caso de dato atípico o valor fuera de rango.

En general, una observación que es inusualmente grande o pequeña en relación con los demás valores de un conjunto de datos se denomina **dato atípico o fuera de rango**.

Estos valores son atribuibles, por lo general, a una de las siguientes causas:

1. El valor ha sido introducido en la base de datos incorrectamente.
2. El valor proviene de una población distinta a la que estamos estudiando.
3. El valor es correcto pero representa un suceso muy poco común.

A continuación vamos a proponer dos maneras de determinar si un dato es un valor fuera de rango.

2.6.1. Mediante la regla empírica

Este método es adecuado si el histograma de los datos tiene forma de campana, en cuyo caso podemos aplicar la regla empírica para detectar qué datos están fuera de los rangos *lógicos* según esta regla.

Según ella, el 99.5 % de los datos están en el intervalo $[\bar{x} - 3s_{n-1}, \bar{x} + 3s_{n-1}]$, luego *se considerarán datos atípicos los x_i que no pertenezcan al intervalo* $[\bar{x} - 3s_{n-1}, \bar{x} + 3s_{n-1}]$.

2.6.2. Mediante los percentiles

Supongamos que tenemos un conjunto de datos $x_1, ..., x_n$. El procedimiento es el siguiente:

1. Se calculan los cuartiles primero y tercero, es decir, los percentiles 25 y 75, P_{25} y P_{75}. Se calcula el llamado *rango intercuartílico* (IR o RI), $IR = P_{75} - P_{25}$.
2. Se consideran **datos atípicos** aquellos inferiores a $P_{25} - 1.5IR$ o superiores a $P_{75} + 1.5IR$.

	Medias	Desv. Típica	CV	Coef. Asimetría
Serie 1	92.01	3.62	25.40	-1.79
Serie 2	92.74	3.73	24.86	1.71

Cuadro 2.2: Resumen descriptivo de los datos de las placas de silicio

Ejemplo. Vamos a ver si hay algún dato atípico entre los datos de tiempo de procesado en una CPU de 25 tareas.

Dado que el histograma no tenía forma de campana, el método de la regla empírica no es el método más adecuado para la detección de valores atípicos.

Por su parte, $P_{50} = 1.38$, $P_{25} = 0.82$ y $P_{75} = 2.16$. Por tanto, $IR = 2.16 - 0.82 = 1.34$, y el intervalo fuera del cúal consideramos valores fuera de rango es $[0.82 - 1.5 \times 1.34, 2.16 + 1.5 \times 1.34] = [-1.19, 4.17]$. De esta forma, el valor 4.75 es un valor fuera de rango.

Hay una versión gráfica de este método para detectar valores atípicos mediante los percentiles: se llama **diagrama de caja o diagrama de cajas y bigotes** o (en inglés) **boxplot**. Este diagrama incluye en un gráfico:

1. El valor de la mediana (o segundo cuartil, Q_2): ese es el centro de la caja.
2. El valor de los percentiles 25 y 75, cuartiles primero y tercero respectivamente (Q_1 y Q_3): son los lados inferior y superior de la caja.
3. El diagrama no representa los límites $P_{25} - 1.5 \times IR$ y $P_{75} + 1.5 \times IR$. En su lugar, señala los últimos puntos no atípicos por debajo (L_i) y por encima (L_s), es decir, señala el último dato por encima de $P_{25} - 1.5 \times IR$ y el último dato por debajo de $P_{75} + 1.5 \times IR$, y los representa como *bigotes* que salen de la caja.
4. Normalmente representa con círculos los datos atípicos.

2.7. Sobre el ejemplo de las capas de dióxido de silicio

Ya estamos en condiciones de responder en parte a las cuestiones que quedaron latentes en el tema de introducción sobre el ejemplo de las placas de silicio.

Vamos a comenzar realizando un resumen descriptivo de los datos, separando por series, proporcionando media, desviación típica, coeficiente de variación y coeficiente de asimetría. Todos estos resultados aparecen en la Tabla 2.2.

En primer lugar, es cierto que, como apuntábamos en el tema de introducción, los valores están en torno a 90 (la media es 92 más o menos). Además, vemos que sí que hay una variabilidad moderada de los datos, con un CV en torno al 25 %, lo que indica que, al parecer, las distintas condiciones en que cada medición se realizó, afectaron en alguna medida el resultado: todo esto es muy preliminar porque no tenemos la información completa de en qué condiciones se realizaron cada una de las mediciones. Por el contrario, podemos observar algo muy llamativo. Los datos de la primera serie son claramente asimétricos a la izquierda (coeficiente de

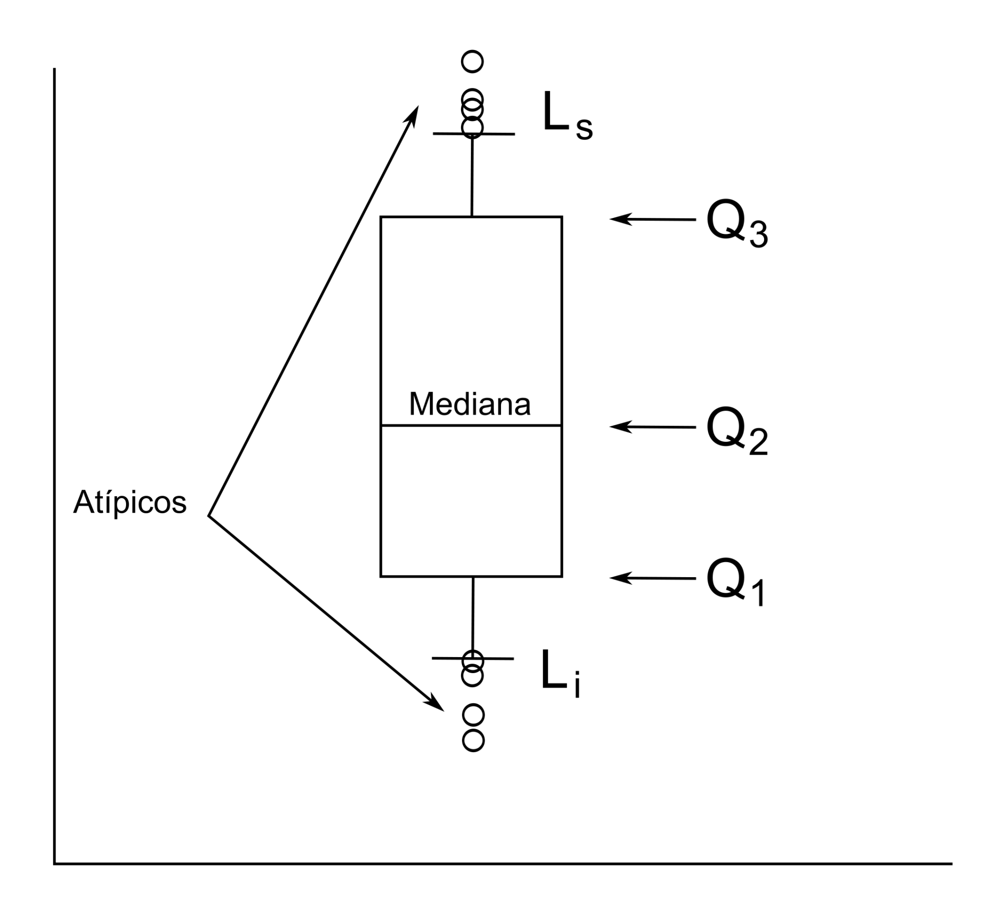


Figura 2.7: Descripción de un diagrama de caja. Fuente: http://es.wikipedia.org/wiki/Diagrama_de_caja

asimetria de -1.79), mientras que los de la segunda serie son claramente asimétricos a la derecha (coeficiente de asimetría de 1.71). Dado que no era esperable que surgieran diferencias entre las dos series, debemos preguntarnos qué pasó.

Para tratar de analizar más profundamente los datos, vamos a proporcionar también los dos diagramas de caja de ambas series. Aparecen en la Figura 2.8. Con ellas, vamos a resumir ahora las decisiones que los autores tomaron en vista de los resultados y las conclusiones a las que llegaron.

Obsérvese que las diferencias entre las series no afectan sorprendentemente al conjunto de las muestras, sino sólo a los valores atípicos que se ven en ambos diagramas de caja. Eso *probaría* que, en efecto, no hay ninguna diferencia sistemática entre las series.

La siguiente tarea es la de inspeccionar los datos atípicos. Si miramos con atención los datos, vemos que las 8 mediciones más grandes de la segunda serie ocurrieron en la placa 10. Al ver este hecho, los autores del trabajo inspeccionaron esta placa y descubrieron que se había contaminado con un residuo de la película, lo que ocasionó esas mediciones tan grandes del espesor. De hecho, los ingenieros eliminaron esa placa y toda la serie entera por razones técnicas. En la primera serie, encontraron también que las tres mediciones más bajas se habían debido a un calibrador mal configurado, por lo que las eliminaron. No se pudo determinar causa alguna a la existencia de los dos datos atípicos restantes, por lo que permanecieron en el análisis. Por último, nótese que después de este proceso de depuración de los datos que el análisis mediante Estadística Descriptiva ha motivado, la distribución de los datos tiene una evidente forma de campana.

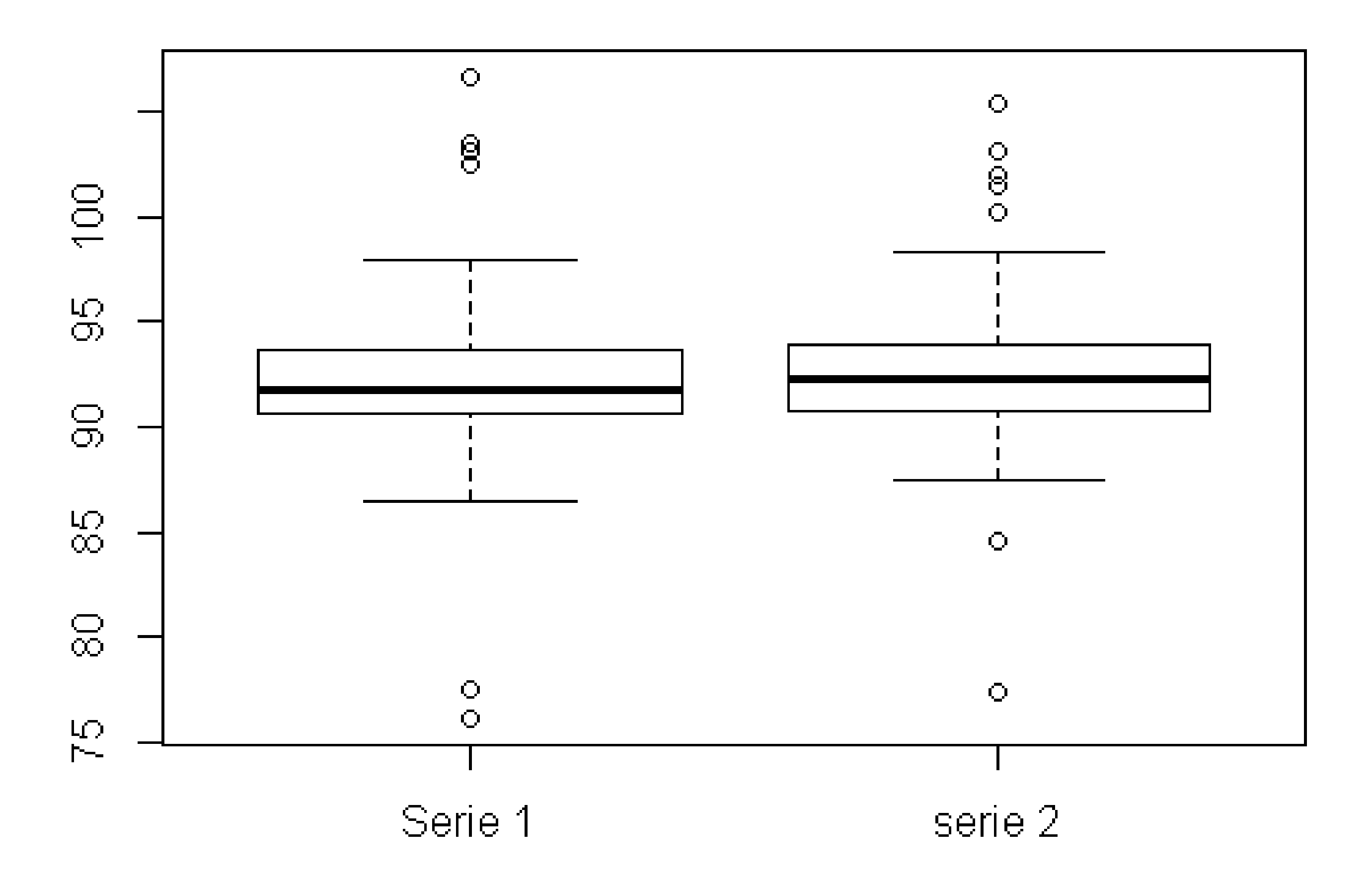


Figura 2.8: Diagramas de caja de los datos del espesor de las capas de dióxido de silicio

Parte II

Cálculo de Probabilidades

Capítulo 3

Probabilidad

> Vemos que la teoría de la probabilidad en el fondo sólo es sentido común reducido a cálculo; nos hace apreciar con exactitud lo que las mentes razonables toman por un tipo de instinto, incluso sin ser capaces de darse cuenta[...] Es sorprendente que esta ciencia, que surgió del análisis de los juegos de azar, llegara a ser el objeto más importante del conocimiento humano[...] Las principales cuestiones de la vida son, en gran medida, meros problemas de probabilidad.
>
> Pierre Simon, Marqués de Laplace

Resumen. El capítulo proporciona un tratamiento de los experimentos cuyos resultados no se pueden predecir con certeza a través del concepto de probabilidad. Se analizan las propiedades de la probabilidad y se introduce también el concepto de probabilidad condicionada, que surge cuando un suceso modifica la asignación de probabilidades previa.

Palabras clave: experimento aleatorio, experimento determinístico, espacio muestral, suceso, probabilidad, probabilidad condicionada, independencia de sucesos.

3.1. Introducción

En nuestra vida cotidiana asociamos usualmente el concepto de **Probabilidad** a su calificativo **probable,** considerando **probables** aquellos eventos en los que tenemos un alto grado de creencia en su ocurrencia. En esta línea, **Probabilidad** es un concepto asociado a la medida del **azar**. También pensamos en el azar vinculado, fundamentalmente, con los juegos de azar, pero desde esa óptica tan reducida se nos escapan otros muchísimos ejemplos de fenómenos de la vida cotidiana o asociados a disciplinas de distintas ciencias donde el azar juega un papel fundamental. Por citar algunos:

- ¿Qué número de unidades de producción salen cada día de una cadena de montaje? No existe un número fijo que pueda ser conocido a priori, sino un conjunto de posibles valores que podrían darse, cada uno de ellos con un cierto grado de certeza.

- ¿Cuál es el tamaño de un paquete de información que se transmite a través de HTTP? No existe en realidad un número fijo, sino que éste es desconocido a priori.

- ¿Cuál es la posición de un objeto detectado mediante GPS? Dicho sistema obtiene, realmente, una estimación de dicha posición, pero existen márgenes de error que determinan una región del plano donde el objeto se encuentra con alta probabilidad.
- ¿Qué ruido se adhiere a una señal que se envía desde un emisor a un receptor? Dependiendo de las características del canal, dicho ruido será más o menos relevante, pero su presencia no podrá ser conocida a priori, y deberá ser diferenciada de la señal primitiva, sin que se conozca ésta, teniendo en cuenta que se trata de un ruido *aleatorio*.

En todos estos ejemplos el azar es un factor insoslayable para conocer el comportamiento del fenómeno en estudio.

3.2. Experimentos aleatorios y experimentos determinísticos

En general, un experimento del que se conocen todos sus posibles resultados y que, repetido en las mismas condiciones, no siempre proporciona los mismos resultados se conoce como **experimento aleatorio**.
En contraposición, un **experimento determinístico** es aquel donde las mismas condiciones aseguran que se obtengan los mismos resultados.

Lo que el Cálculo de Probabilidades busca es encontrar una medida de la incertidumbre o de la certidumbre que se tiene de todos los posibles resultados, ya que jamás (o muy difícilmente) se podrá conocer a priori el resultado de cualquier experimento donde el azar esté presente: a esta medida de la incertidumbre la denominaremos *probabilidad*[1].

3.3. Definición de probabilidad

Tenemos, por tanto, que probabilidad es la asignación que hacemos del grado de creencia que tenemos sobre la ocurrencia de algo. Esta asignación, sin embargo, debe ser *coherente*. Esta necesidad de que asignemos probabilidades adecuadamente se va a plasmar en esta sección en tres reglas, conocidas como *axiomas*, que debe cumplir cualquier reparto de probabilidades.

3.3.1. Álgebra de conjuntos

Si consideramos un experimento aleatorio, podemos caracterizar los posibles resultados de dicho experimento como conjuntos. Es de interés, por tanto, repasar los conceptos y propiedades básicas del álgebra de conjuntos.

En todo este apartado no debemos olvidar que los conjuntos representan en nuestro caso los posibles resultados de un experimento aleatorio.

Un **conjunto** es una colección de elementos.
Se dice que B es un **subconjunto de** A si todos sus elementos lo son también de A, y se notará $B \subset A$.

[1] Es mejor que aceptemos desde el principio que la Estadística no es la ciencia de la adivinación: tan sólo se ocupa de cuantificar cómo de incierto es un evento y, ocasionalmente, de proponer estrategias de predicción basadas en dicha medida de la incertidumbre.

Para cada A se verifica $\emptyset \subset A \subset A \subset \Omega$.

Si $C \subset B$ y $B \subset A$, entonces, $C \subset A$. Esto se conoce como propiedad transitiva.

La **unión** de B y A es un conjunto cuyos elementos son los elementos de A y B, y se nota $A \cup B$. Esta operación verifica la propiedad conmutativa y asociativa.

Si $A \subset B$, entonces $A \cup B = B$.

La **intersección** de A y B es el conjunto formado por los elementos comunes de A y B, y se nota AB o $A \cap B$. Esta operación verifica la propiedad conmutativa, asociativa y distributiva respecto de la unión.
Dos conjuntos, A y B, se dicen **mutuamente excluyentes, disjuntos o incompatibles** si su intersección es vacía, es decir, $A \cap B = \emptyset$.

Si dos conjuntos A y B son disjuntos, su unión suele notarse $A + B$.

Los conjuntos $A_1, ..., A_N$ se dicen **mutuamente excluyentes** si $A_i \cap A_j = \emptyset$ para todo $i \neq j$.
Una **partición** es una colección de conjuntos, $A_1, ..., A_N$ tal que:
a) $A_1 \cup ... \cup A_N = \Omega$
b) $A_i \cap A_j = \emptyset$ para todo $i \neq j$.
El **conjunto complementario** de un conjunto A, $\bar{A}$ ó A^c, está formado por todos los elementos de Ω que no pertenecen a A.

Se sigue por tanto,

$$
\begin{aligned}
A \cup \bar{A} &= \Omega \\
A \cap \bar{A} &= \emptyset \\
(A^c)^c &= A \\
\bar{\Omega} &= \emptyset \\
\text{Si } B \subset A &\rightarrow \bar{A} \subset \bar{B} \\
\text{Si } A = B &\rightarrow \bar{A} = \bar{B}.
\end{aligned}
$$

Finalmente, mencionemos las llamadas Leyes de Morgan:

$$\overline{A \cup B} = \bar{A} \cap \bar{B}$$

$$\overline{A \cap B} = \bar{A} \cup \bar{B}.$$

3.3.2. Espacio muestral

Consideremos un experimento aleatorio.

El conjunto formado por todos los posibles resultados del experimento aleatorio recibe el nombre de **espacio muestral**, y lo notaremos habitualmente como Ω.
Cualquier subconjunto de un espacio muestral recibe el nombre de **suceso** o **evento**.
Hablaremos de **ensayo o realización** de un experimento aleatorio refiriéndonos a una ejecución de dicho experimento. Así, diremos que en un ensayo **ocurre un suceso** A si se observa en dicho ensayo cualquier resultado incluido en el suceso A.

Una observación importante es que el espacio muestral no tiene por qué ser único, sino que dependerá de lo que deseemos observar del experimento aleatorio. Vamos a poner este hecho de manifiesto en los siguientes ejemplos.

Ejemplo. Si consideramos el lanzamiento de un dado, un espacio muestral sería Ω={1,2,3,4,5,6}.

Los sucesos más elementales posibles son {1}, {2}, {3}, {4}, {5} y {6}. Otros sucesos no elementales pueden ser {1,2}, {mayor que 2}, {par}, ...

Sin embargo, supongamos que estamos lanzando un dado porque no tenemos ninguna moneda a mano, y sólo deseamos ver si el resultado es par o impar. En ese caso, el espacio muestral sería $\Omega = \{par, impar\}$.

Ejemplo. Un experimento habitual en Biología consiste en extraer, por ejemplo, peces de un río, hasta dar con un pez de una especie que se desea estudiar. El número de peces que habría que extraer hasta conseguir el ejemplar deseado de la especie en estudio formaría el espacio muestral, $\Omega = \{1, 2, 3, ...\}$, si es que el investigador desea observar exactamente el número de peces hasta extraer ese ejemplar deseado. Obsérvese que se trata de un conjunto no acotado, pero numerable.

Como ejemplos de posibles sucesos de interés podríamos poner los eventos {1,2,3,4,5}, {mayor o igual a 5},...

Supongamos ahora que el investigador sólo está interesado en comprobar si hacen falta más de 5 extracciones para obtener un ejemplar de la especie en estudio. En ese caso, el espacio muestral sería $\Omega = \{> 5, \leq 5\}$.

Ejemplo. Si consideramos el experimento aleatorio consistente en elegir un número absolutamente al azar entre 0 y 1, un espacio muestral sería $\Omega = [0, 1]$. A diferencia de los anteriores ejemplos, este espacio muestral no es finito, ni siquiera numerable.

Como ejemplo de sucesos posibles en este espacio muestral podemos destacar, entre otros, {menor que 0.5} , {mayor que 0.25}, {menor que 0.75} ,...

Otro espacio muestral podría ser observar el valor decimal mayor más cercano. Por ejemplo, si sale 0.25, me interesa 0.3. En ese caso el espacio muestral sería $\Omega = 0.1, 0.2, ...1$. Este espacio muestral serviría, por ejemplo, para sortear números entre 1 y 10, sin más que multiplicar el resultado obtenido por 10.

En estos últimos ejemplos podemos ver que hay dos grandes tipos de espacios muestrales según el número de sucesos elementales.

Un espacio muestral se dice **discreto** si está formado por un conjunto finito o infinito numerable de sucesos elementales.
Por el contrario, un espacio muestral se dice **continuo** si está formado por un conjunto no numerable de sucesos elementales.

3.3.3. Función de probabilidad

Dado un espacio muestral Ω correspondiente a un experimento aleatorio, una **función de probabilidad** para ese espacio muestral es cualquier función que asigne a cada suceso un número en el intervalo $[0,1]$ y que verifique
$P[A] \geq 0$, para cualquier evento A.
$P[\Omega] = 1$.
Dada una colección de sucesos $A_1, A_2, ..., A_n$ mutuamente excluyentes, es decir, tales que $A_i \cap A_j = \emptyset$ para todo $i \neq j$,

$$P[\cup_{i=1}^{n} A_i] = \sum_{i=1}^{n} P[A_i].$$

Nota. Hay que notar que se puede dar más de una función de probabilidad asociada al mismo espacio muestral. Por ejemplo, asociado al espacio muestral $\Omega = \{cara, cruz\}$, del lanzamiento de una moneda, pueden darse un número infinito no numerable de medidas de la probabilidad; concretamente, asociadas a cada elección

$$P[cara] = p$$
$$P[cruz] = 1 - p,$$

para cada $p \in [0,1]$. Aunque si la moneda no está cargada, como sucede habitualmente, se considera el caso en que $p = \frac{1}{2}$.

Ejemplo. Volviendo sobre el lanzamiento del dado, si éste no está cargado, podemos definir la siguiente función de probabilidad:

$$P[\{i\}] = \frac{1}{6},\ i = 1, 2, ..., 6.$$

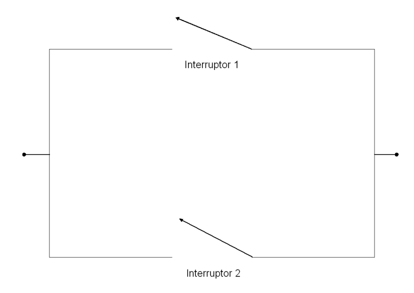


Figura 3.1: Circuito

En ese caso, podemos, a su vez, calcular algunas probabilidades. Por ejemplo,

$$\begin{aligned} P(\{par\}) &= P[\{2,4,6\}] \\ &= P[\{2\}] + P[\{4\}] + P[\{6\}] \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 0.5. \end{aligned}$$

En este cálculo se ha tenido en cuenta la tercera condición de la definición axiomática.

Como consecuencia de la definición se verifican, entre otras, las siguientes propiedades, que además facilitan bastante los cálculos:

- $P[\emptyset] = 0$.
- Sea A un suceso cualquiera. Entonces, $P\left[\bar{A}\right] = 1 - P[A]$.
- Sean A y B dos sucesos cualesquiera. Entonces, $P\left[A \cap \bar{B}\right] = P[A] - P[A \cap B]$.
- Sean A y B dos sucesos cualesquiera. Entonces, $P[A \cup B] = P[A] + P[B] - P[A \cap B]$.

Ejemplo. El circuito que aparece en la Figura 3.1 está constituido por dos interruptores (*switches*) en paralelo. La probabilidad de que cualquiera de ellos esté cerrado es de $\frac{1}{2}$.

Para que pase corriente a través del circuito basta con que pase corriente por alguno de los dos interruptores, esto es, que al menos uno de ellos esté cerrado. Por tanto, si notamos por E al suceso *que pase corriente a través del circuito* y E_i al suceso *que el interruptor i esté cerrado*, entonces,

$$\begin{aligned} P[E] &= P[E_1 \cup E_2] = P[E_1] + P[E_2] - P[E_1 \cap E_2] \\ &= \frac{1}{2} + \frac{1}{2} - P[E_1 \cap E_2] \leq 1. \end{aligned}$$

Para conocer esta probabilidad de forma exacta necesitamos saber cómo actúan de forma conjunta ambos circuitos.

Nº de lanzamientos	10	100	250	500	750	1000
Nº de caras	4	46	124	244	379	501
$\frac{N.\ de\ caras}{N.\ de\ lanzamientos}$	0.4	0.46	0.496	0.488	0.5053	0.501

Cuadro 3.1: Aproximación frecuentista a la probabilidad de cara en el lanzamiento de una moneda.

3.4. Interpretación frecuentista de la probabilidad

La interpretación más común al concepto de probabilidad tiene que ver con los promedios de ocurrencia de los sucesos del experimento en cuestión.

Pensemos en el lanzamiento de una moneda: si decimos que la probabilidad de cara es 0.5, entendemos que si lanzamos la moneda un gran número de veces y anotamos el número de caras, éstas serán más o menos la mitad.

Generalizando este proceso, podríamos decir que la probabilidad de un evento A, $P[A]$, es

$$P[A] = \lim_{n\to\infty} \frac{n_A}{n},$$

donde n_A es el número de ocurrencias de A en n ensayos del experimento.

Esta interpretación se conoce como *definición frecuentista de la probabilidad*. Se trata de una interpretación de carácter eminentemente práctico porque permite una aproximación física al concepto de probabilidad, pero se ve limitada por las complicaciones que supone la definición en términos de un límite que, como tal, sólo se alcanza *en el infinito*. Además, desde un punto de vista realista, ¿en qué ocasiones podremos repetir el experimento un gran número de veces?

Ejemplo. Se han realizado 1000 lanzamientos de una moneda. En el Cuadro 3.1 aparece un resumen de ese proceso. Puede observarse como cuanto mayor es el número de lanzamientos, más se aproxima la frecuencia relativa al valor $\frac{1}{2}$, de manera que podríamos pensar que la probabilidad de cara es igual que la probabilidad de cruz e iguales ambas a $\frac{1}{2}$, aunque esto sólo es una suposición, o una aproximación, ya que para aplicar estrictamente la definición frecuentista deberíamos continuar hasta el infinito, lo que resulta imposible.

Esta interpretación frecuentista de la probabilidad permite inferir lo que podemos llamar *frecuencias esperadas*. Si un evento A tiene asignada una probabilidad $P[A]$, entonces, si repetimos el experimento aleatorio n veces, *lo más esperable* es que el número de veces que se de el evento A será $n \times P[A]$. Más adelante podremos matizar con más rigor a qué nos referimos con *lo más esperable*.

Ejemplo. Siguiendo con el ejemplo de la moneda, si la lanzamos 348 veces, lo esperable es que salgan alrededor de $348 \times 0.5 = 174$ caras.

3.5. Interpretación subjetiva de la probabilidad

Si nos dicen que la probabilidad de que llueva mañana es del 35 %, ¿cómo podemos interpretar eso en términos frecuentistas? No tiene sentido pensar en que podemos repetir el experimento *día de mañana* muchas veces y contar cuántas veces llueve. ¿Podríamos pensar *si hubiera muchos días como el de mañana, aproximadamente llovería en el 35 % de ellos*? Pero eso no tiene sentido porque el día de mañana es único.

La interpretación subjetiva de la probabilidad tiene que ver con la vinculación de este concepto con el grado de incertidumbre que tenemos sobre las cosas. Si tenemos un experimento aleatorio, el resultado de dicho experimento es incierto. La probabilidad de un resultado del experimento es el grado de creencia que yo tengo en la ocurrencia de dicho resultado. Ese grado de creencia es personal, luego es subjetivo, pero lógicamente, deberá estar acorde con la información que tenemos sobre el experimento.

3.6. Espacio muestral con resultados equiprobables. Fórmula de Laplace

Otro punto de vista que permite abordar el proceso de asignación de probabilidad a sucesos es el siguiente: continuando con el ejemplo de la moneda, en este experimento son dos los resultados posibles, y no hay razones para pensar que uno de ellos es *más probable* que otro, así que tiene sentido considerar que la probabilidad de cara y la probabilidad de cruz son ambas del 50 %.

En general, si el espacio muestral está formado por N resultados posibles y todos ellos tienen la misma probabilidad (equiprobables), podríamos decir que la probabilidad de un evento A, $P[A]$, es

$$P[A] = \frac{N_A}{N},$$

donde N_A es el número de resultados favorables a la ocurrencia de A.

Esta fórmula, conocida como *fórmula de Laplace* también es fundamentalmente práctica. Por ejemplo, nos permite deducir que

$$P[cara] = \frac{1}{2}$$

en el lanzamiento de una moneda sin tener que lanzar la moneda un gran número de veces.

Sin embargo, la definición tiene dos grandes inconvenientes: el conjunto de resultados posibles, N, tiene que ser finito y, además, todos los resultados posibles deben tener la misma probabilidad (con lo cual, lo definido queda implícitamente inmerso en la definición).

3.7. Probabilidad condicionada. Independencia de sucesos

Para introducir de manera intuitiva el concepto de probabilidad condicionada debemos pensar en la probabilidad como medida de la creencia en la ocurrencia de los sucesos.

Pensemos en un experimento aleatorio y en un suceso de dicho experimento, A, en el que, en principio, tenemos un grado de creencia $P[A]$; pero supongamos que conocemos algo del resultado de dicho experimento; concretamente, sabemos que ha ocurrido un suceso B. Parece lógico pensar que esa información conocida sobre el resultado del ensayo modificará nuestro grado de creencia en A: llamemos a este nuevo grado de creencia $P[A \mid B]$, **probabilidad de A conocida B** o **probabilidad de A condicionada a B**.

Ejemplo. Consideremos el suceso A : el día de hoy va a llover y el suceso B : el día de hoy está nublado. Obviamente, la probabilidad $P[A]$ será menor que la probabilidad $P[A \mid B]$, ya que el hecho de que esté nublado refuerza nuestra creencia en que llueva.

Ejemplo. Consideremos el experimento aleatorio de extraer una carta de una baraja española. Sea el suceso A : obtener una sota, el suceso B_1 : obtener una figura y el suceso B_2 : obtener una carta de copas.
Las distintas probabilidades, condicionadas o no, bajo la definición clásica, son las siguientes:

$$\begin{aligned} P[A] &= \frac{4\ sotas}{40\ cartas} = \frac{1}{10} \\ P[A \mid B_1] &= \frac{4\ sotas}{12\ figuras} = \frac{1}{3} \\ P[A \mid B_2] &= \frac{1\ sota\,de\,copas}{10\,copas} = \frac{1}{10}. \end{aligned}$$

Como puede verse, B_1 modifica la probabilidad a priori, pero no así B_2. Puede decirse que B_2 no ofrece información acerca de A, o que A y B_2 son **independientes**.

Vamos a dar a continuación una definición de **probabilidad condicionada** que responde a esta idea de recalcular la probabilidad en función de la información existente.

La **probabilidad condicionada de un suceso** A**, conocido otro suceso** B, denotada por $P[A \mid B]$, se define como el cociente

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]},$$

siempre que $P[B] \neq 0$.

Una función de probabilidad condicionada $P[\cdot/_B]$ es una función de probabilidad en toda regla: por tanto, cumple las mismas propiedades que cualquier función de probabilidad "sin condicionar".

Como hemos comentado, la idea de la probabilidad condicionada es utilizar la información que nos da un suceso conocido sobre la ocurrencia de otro suceso. Pero, como ya hemos puesto de manifiesto en un ejemplo, no siempre un suceso da información sobre otro. En este caso se dice que ambos sucesos son **independientes**. Por tanto:

Dos sucesos A y B se dicen independientes si $P[A \mid B] = P[A]$, o equivalentemente si $P[B \mid A] = P[B]$, o equivalentemente si $P[A \cap B] = P[A] \times P[B]$.

Ejemplo. Continuando con el Ejemplo 3.3.3, lo más lógico es pensar que los dos interruptores actúan de forma independiente, en cuyo caso $P[E_1 \cap E_2] = P[E_1]\,P[E_2]$ y tenemos que,

$$\begin{aligned} P[E] &= \frac{1}{2} + \frac{1}{2} - P[E_1 \cap E_1] \\ &= \frac{1}{2} + \frac{1}{2} - \frac{1}{2}\frac{1}{2} = \frac{3}{4}. \end{aligned}$$

Nota. Es muy importante no confundir la probabilidad condicionada de un suceso a otro con la probabilidad de la intersección de ambos sucesos. En la Figura 3.2 puede verse la diferencia entre las probabilidades condicionadas entre dos sucesos y la probabilidad de su intersección. En términos coloquiales, podemos

analizar estas probabilidades como el cociente entre *una parte* y *un todo*. Cuando la probabilidad es condicionada ese *todo* es el suceso que condiciona. Cuando la probabilidad no es condicionada, ese *todo* es todo el espacio muestral. En ambos casos esa *parte* es la intersección.

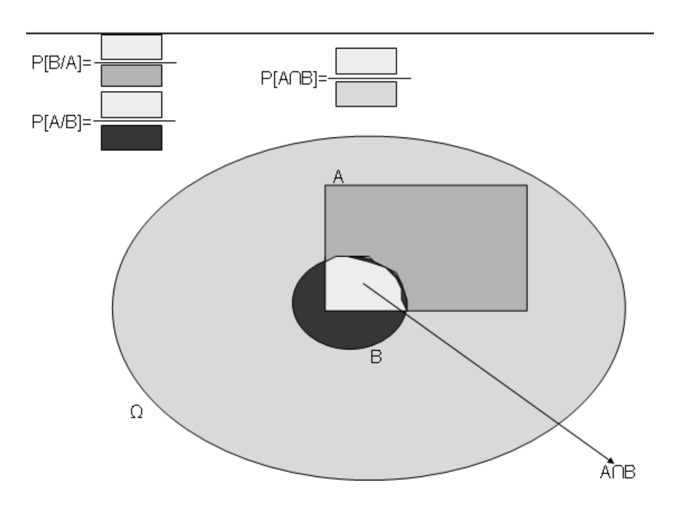


Figura 3.2: Esquema acerca de la definición de probabilidad condicionada.

Nota. También suele ser bastante común la confusión entre sucesos independientes y sucesos incompatibles o mutuamente excluyentes.

En este sentido, recordemos que dos sucesos A y B son incompatibles o mutuamente excluyentes si $A \cap B = \varnothing$, en cuyo caso $P[A \cap B] = 0$.

Por su parte, A y B serán independientes si $P[A \cap B] = P[A]\,P[B]$.

Las diferencias entre ambos conceptos son obvias.

Ejemplo. La probabilidad de que el producto no sea elaborado a tiempo es 0.05. Se solicitan tres pedidos del producto con la suficiente separación en el tiempo como para considerarlos eventos independientes.

1. ¿Cuál es la probabilidad de que todos los pedidos se envíen a tiempo?

   En primer lugar, notemos E_i al suceso *enviar a tiempo el pedido i-ésimo*. En ese caso, sabemos que $P[E_i] = 0.95$.

   Por su parte, nos piden

   $$P[E_1 \cap E_2 \cap E_3] = P[E_1]\,P[E_2]\,P[E_3] = 0.95^3,$$

   debido a que los pedidos son independientes.

2. ¿Cuál es la probabilidad de que exactamente un pedido no se envíe a tiempo?

En este caso el suceso que nos piden es más complejo:

$$P\left[\bar{E}_1 \cap E_2 \cap E_3 \cup E_1 \cap \bar{E}_2 \cap E_3 \cup E_1 \cap E_2 \cap \bar{E}_3\right]$$

$$= P\left[\bar{E}_1 \cap E_2 \cap E_3\right] + P\left[E_1 \cap \bar{E}_2 \cap E_3\right] + P\left[E_1 \cap E_2 \cap \bar{E}_3\right]$$

$$= 0.05 \times 0.95^2 + 0.05 \times 0.95^2 + 0.05 \times 0.95^2 = 0.135,$$

donde se ha utilizado que los sucesos $\bar{E}_1 \cap E_2 \cap E_3$, $E_1 \cap \bar{E}_2 \cap E_3$ y $E_1 \cap E_2 \cap \bar{E}_3$ son incompatibles.

3. ¿Cuál es la probabilidad de que dos o más pedidos no se envíen a tiempo?

   Tengamos en cuenta que ya hemos calculado la probabilidad de que todos se envíen a tiempo y de que todos menos uno se envíen a tiempo. Entonces,

$$P\left[\text{dos o más pedidos no se envíen a tiempo}\right]$$

$$= 1 - P\left[\text{todos se envíen a tiempo} \cup \text{un pedido no se envíe a tiempo}\right]$$
$$= 1 - \left(0.95^3 + 0.135\right).$$

Ejemplo. Consideremos un proceso industrial como el que se esquematiza en la Figura 3.3. En dicho esquema se pone de manifiesto que una unidad será producidad con éxito si pasa en primer lugar un chequeo previo (A); después puede ser montada directamente (B), redimensionada (C) y después montada (D) o adaptada (E) y después montada (F); posteriormente debe ser pintada (G) y finalmente embalada (H). Consideremos que las probabilidades de pasar exitosamente cada subproceso son todas ellas iguales a 0.95, y que los subprocesos tienen lugar de forma independiente unos de otros. Vamos a calcular en esas condiciones la probabilidad de que una unidad sea exitosamente producida.

Si nos damos cuenta, A, G y H son ineludibles, mientras que una unidad puede ser producida si pasa por B, por C y D o por E y F. En notación de conjuntos, la unidad será producida si se da

$$A \cap (B \cup C \cap D \cup E \cap F) \cap G \cap H.$$

Como los procesos son independientes unos de otros, no tenemos problemas con las probabilidades de las intersecciones, pero tenemos que calcular la probabilidad de una unión de tres conjuntos, $B \cup C \cap D \cup E \cap F$. En general,

$$P\left[A_1 \cup A_2 \cup A_3\right] = P\left[(A_1 \cup A_2) \cup A_3\right] = P\left[A_1 \cup A_2\right] + P\left[A_3\right] - P\left[(A_1 \cup A_2) \cap A_3\right]$$
$$= P\left[A_1\right] + P\left[A_2\right] - P\left[A_1 \cap A_2\right] + P\left[A_3\right] - P\left[A_1 \cap A_3 \cup A_2 \cap A_3\right]$$

$$
\begin{aligned}
&= P[A_1] + P[A_2] - P[A_1 \cap A_2] + P[A_3] \\
&- (P[A_1 \cap A_3] + P[A_2 \cap A_3] - P[A_1 \cap A_2 \cap A_3])
\end{aligned}
$$

$$
\begin{aligned}
&= P[A_1] + P[A_2] + P[A_3] \\
&- P[A_1 \cap A_2] - P[A_1 \cap A_3] - P[A_2 \cap A_3] \\
&+ P[A_1 \cap A_2 \cap A_3]
\end{aligned}
$$

En nuestro caso,

$$
\begin{aligned}
P[B \cup C \cap D \cup E \cap F] &= P[B] + P[C \cap D] + P[E \cap F] \\
&- P[B \cap C \cap D] - P[B \cap E \cap F] - P[C \cap D \cap E \cap F] \\
&+ P[B \cap C \cap D \cap E \cap F] \\
&= 0.95 + 2 \times 0.95^2 - 2{\times}0.95^3 - 0.95^4 + 0.95^5 \\
&= 0.9995247
\end{aligned}
$$

Ya estamos en condiciones de obtener la probabilidad que se nos pide:

$$
\begin{aligned}
P[A \cap (B \cup C \cap D \cup E \cap F) \cap G \cap H] &= P[A]\, P[B \cup C \cap D \cup E \cap F]\, P[G]\, P[H] \\
&= 0.95 \times (0.9995247) \times 0.95 \times 0.95 \\
&= 0.8569675.
\end{aligned}
$$

En estos ejemplos, el cálculo de la probabilidad de las intersecciones ha resultado trivial porque los sucesos son independientes. Son embargo, esto no siempre ocurre. ¿Cómo podemos, en general, obtener la probabilidad de la intersección de dos o más sucesos no necesariamente independientes?

En el caso de sólo dos sucesos, A y B, podemos deducir que

$$P[A \cap B] = P[A|B] \times P[B]$$

directamente de la definición de probabilidad condicionada. A partir de esta fórmula, por inducción, se puede obtener la llamada fórmula producto, que se enuncia de la siguiente forma: si $A_1, A_2, ..., A_n$ son sucesos de un espacio muestral no necesariamente independientes, se verifica

$$P[A_1 \cap A_2 \cap ... \cap A_n] = P[A_1]P[A_2|A_1]...P[A_n|A_1 \cap A_2 \cap ... \cap A_{n-1}]$$

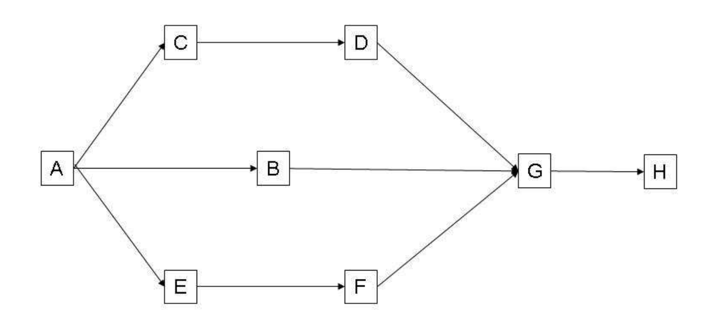


Figura 3.3: Esquema del proceso industrial del ejemplo

Ejemplo. Un lote de 50 arandelas contiene 30 arandelas cuyo grosor excede las especificaciones de diseño. Suponga que se seleccionan 3 arandelas al azar y sin reemplazo del lote.

1. ¿Cuál es la probabilidad de que las tres arandelas seleccionadas sean más gruesas que las especificaciones de diseño?

   Comenzamos notando los sucesos A_i: la í-ésima arandela extraida es más gruesa que las especificaciones de diseño, $i = 1, 2, 3$.

   Entonces, nos piden

$$\begin{aligned} P\left[A_1 \cap A_2 \cap A_3\right] &= P\left[A_1\right] P\left[A_2/_{A_1}\right] P\left[A_3/_{A_1 \cap A_2}\right] \\ &= \frac{30}{50} \frac{29}{49} \frac{28}{48}. \end{aligned}$$

2. ¿Cuál es la probabilidad de que la tercera arandela seleccionada sea más gruesa que las especificaciones de diseño si las dos primeras fueron más delgadas que la especificación?

$$P\left[A_3/_{\bar{A}_1 \cap \bar{A}_2}\right] = \frac{30}{48}.$$

3.8. Teorema de la probabilidad total y Teorema de Bayes

Los siguientes dos resultados se conocen como **Teorema de la probabilidad total y Teorema de Bayes** respectivamente, y juegan un importante papel a la hora de calcular probabilidades. Los dos utilizan como

principal herramienta el concepto de probabilidad condicionada.

Teorema de la Probabilidad Total. Sea P una función de probabilidad en un espacio muestral. Sea $\{A_1, ..., A_N\} \subset F$ una partición del espacio muestral Ω y sea B un suceso cualquiera. Entonces,

$$P[B] = P[B \mid A_1] P[A_1] + ... + P[B \mid A_N] P[A_N].$$

Teorema de Bayes. En esas mismas condiciones, si $P[B] \neq 0$,

$$P[A_i \mid B] = \frac{P[B \mid A_i] P[A_i]}{P[B \mid A_1] P[A_1] + ... + P[B \mid A_N] P[A_N]}.$$

Ejemplo. Supongamos que tenemos 4 cajas con componentes electrónicas dentro. La caja 1 contiene 2000 componentes, con un 5 % de defectuosas; la caja 2 contiene 500 componentes, con un 40 % de defectuosas; las cajas 3 y 4 contienen 1000 componentes, con un 10 % de defectuosas.

1. ¿Cuál es la probabilidad de escoger al azar una componente defectuosa?

   Notemos D : componente defectuosa y C_i : componente de la caja i-ésima. Entonces, se tiene que

$$P[C_1] = \frac{2000}{2000+500+1000+1000} = \frac{4}{9}$$
$$P[C_2] = \frac{500}{2000+500+1000+1000} = \frac{1}{9}$$
$$P[C_3] = \frac{1000}{2000+500+1000+1000} = \frac{2}{9}$$
$$P[C_4] = \frac{1000}{2000+500+1000+1000} = \frac{2}{9}$$

   Además, $P[D \mid C_1] = 0.05$, $P[D \mid C_2] = 0.4$, $P[D \mid C_3] = 0.1$ y $P[D \mid C_4] = 0.1$.

   Utilizando el Teorema de la probabilidad total,

$$\begin{aligned} P[D] &= P[D \mid C_1] P[C_1] + P[D \mid C_2] P[C_2] + P[D \mid C_3] P[C_3] \\ &+ P[D \mid C_4] P[C_4] \\ &= 0.05\frac{4}{9} + 0.4\frac{1}{9} + 0.1\frac{2}{9} + 0.1\frac{2}{9} = 0.11111 \end{aligned}$$

2. Si se escoge una componente al azar y resulta ser defectuosa, ¿cuál es la probabilidad de que pertenezca a la caja 1?

$$P[C_1 \mid D] = \frac{P[D \mid C_1] P[C_1]}{P[D]} = \frac{0.05\frac{4}{9}}{0.11111} = 0.2$$

	Número	en cada	caja	
μF	1	2	3	Total
0.01	20	95	25	140
0.1	55	35	75	165
1.0	70	80	145	295
Total	145	210	245	600

Cuadro 3.2: Acumuladores.

Ejemplo. Se disponen tres cajas donde se almacenan acumuladores según aparece en el Cuadro 3.2. Se escoge al azar una caja y de ella, a su vez, un acumulador.

1. ¿Cuál es la probabilidad de que se haya seleccionado un acumulador de $0.01\mu F$?

   Notemos $0.01\mu F$, $0.1\mu F$ y $1.0\mu F$ a los sucesos *extraer un acumulador de* $0.01\mu F$, $0.1\mu F$ y $1.0\mu F$ respectivamente. De igual forma, notemos $c1$, $c2$ y $c3$ a los sucesos *elegir la caja 1, la caja 2 y la caja 3*, respectivamente. Utilizando el teorema de la probabilidad total,

$$\begin{aligned} P[0.01\mu F] &= P[0.01\mu F\,/\,c1]\,P[c1] + P[0.01\mu F\,/\,c2]\,P[c2] + P[0.01\mu F\,/\,c3]\,P[c3] \\ &= \frac{20}{145}\frac{1}{3} + \frac{95}{210}\frac{1}{3} + \frac{25}{245}\frac{1}{3} = \frac{5903}{25\,578} = 0.23078. \end{aligned}$$

2. Si ha sido seleccionado un acumulador de $1.0\mu F$, ¿cuál es la probabilidad de que proceda de la caja 1? Utilizando el teorema de Bayes,

$$P[c1\,/\,1.0\mu F] = \frac{P[1.0\mu F\,/\,c1]\,P[c1]}{P[1.0\mu F]}.$$

   Por su parte,

$$\begin{aligned} P[1.0\mu F] &= P[1.0\mu F\,/\,c1]\,P[c1] + P[1.0\mu F\,/\,c2]\,P[c2] + P[1.0\mu F\,/\,c3]\,P[c3] \\ &= \frac{70}{145}\frac{1}{3} + \frac{80}{210}\frac{1}{3} + \frac{145}{245}\frac{1}{3} = \frac{6205}{12\,789} = 0.48518, \end{aligned}$$

   luego

$$P[c1\,/\,1.0\mu F] = \frac{\frac{70}{145}\frac{1}{3}}{\frac{6205}{12\,789}} = \frac{2058}{6205} = 0.33167.$$

Ejemplo. Siguiendo con el ejemplo de las arandelas con grosor fuera de las especificaciones de diseño, ¿cuál es la probabilidad de que la tercera arandela seleccionada sea más gruesa que las especificaciones de diseño?

$$P[A_3] = P[A_3|_{A_1\cap A_2}]P[A_1\cap A_2] + P[A_3|_{\bar{A}_1\cap A_2}]P[\bar{A}_1\cap A_2]$$

$$+P[A_3|_{A_1\cap \bar{A}_2}]P[A_1\cap \bar{A}_2] + P[A_3|_{\bar{A}_1\cap \bar{A}_2}]P[\bar{A}_1\cap \bar{A}_2]$$

$$= P[A_3|_{A_1\cap A_2}]P[A_1]P[A_2|_{A_1}] + P[A_3|_{\bar{A}_1\cap A_2}]P[\bar{A}_1]P[A_2|_{\bar{A}_1}]$$

$$+P[A_3|_{A_1\cap \bar{A}_2}]P[A_1]P[\bar{A}_2|_{A_1}] + P[A_3|_{\bar{A}_1\cap \bar{A}_2}]P[\bar{A}_1]P[\bar{A}_2|_{\bar{A}_1}]$$

$$= \frac{28}{48}\frac{30}{50}\frac{29}{49} + \frac{29}{48}\frac{20}{50}\frac{30}{49}$$

$$+\frac{29}{48}\frac{30}{50}\frac{20}{49} + \frac{30}{48}\frac{20}{50}\frac{19}{49}.$$

Ejemplo. En el canal de comunicaciones ternario que se describe en la Figura 3.4, se ha observado que el dígito 3 es enviado tres veces más frecuentemente que 1, y 2 dos veces más frecuentemente que 1. Calculemos la probabilidad de que un dígito cualquiera enviado a través del canal sea recibido correctamente.

En primer lugar, si notamos $P[X=1]=p$, entonces $P[X=2]=2p$ y $P[X=3]=3p$. Por otra parte, como

$$1 = P[X=1] + P[X=2] + P[X=3] = 6p,$$

se tiene que

$$P[X=1] = \frac{1}{6},\ P[X=2] = \frac{1}{3} \text{ y } P[X=3] = \frac{1}{2}.$$

Ahora, utilizando el teorema de la probabilidad total,

$$\begin{aligned} P[\textit{dígito}\, OK] &= P[\textit{dígito}\, OK \,/\, X=1]\, P[X=1] \\ &+ P[\textit{dígito}\, OK \,/\, X=2]\, P[X=2] \\ &+ P[\textit{dígito}\, OK \,/\, X=3]\, P[X=3] \\ &= P[Y=1\,/\,X=1]\, P[X=1] \\ &+ P[Y=2\,/\,X=2]\, P[X=2] \\ &+ P[Y=3\,/\,X=3]\, P[X=3] \\ &= (1-\alpha)\frac{1}{6} + (1-\beta)\frac{1}{3} + (1-\gamma)\frac{1}{2} = P. \end{aligned}$$

Ejemplo. Continuando con el anterior, si se recibe un 1, ¿cuál es la probabilidad de que se hubiera enviado un 1?

Utilizando el teorema de Bayes,

$$P[X=1\,/\,Y=1] = \frac{P[Y=1\,/\,X=1]\, P[X=1]}{P[Y=1]}.$$

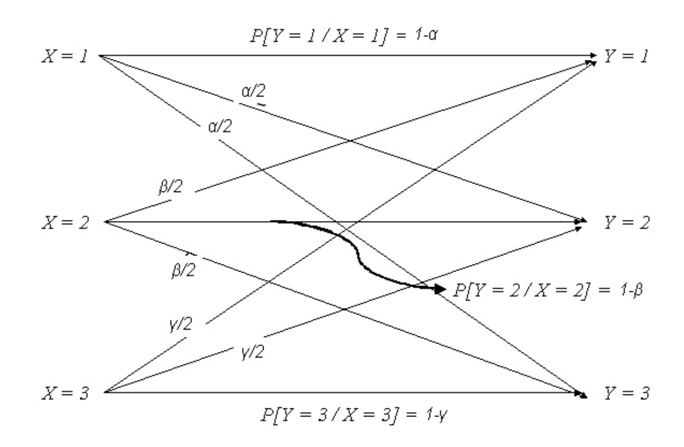


Figura 3.4: Canal ternario de comunicaciones con probabilidad de cruce

Por su parte,

$$\begin{aligned} P[Y=1] &= P[Y=1 / X=1] P[X=1] \\ &+ P[Y=1 / X=2] P[X=2] \\ &+ P[Y=1 / X=3] P[X=3] \\ &= \frac{1-\alpha}{6} + \frac{\beta}{6} + \frac{\gamma}{4}, \end{aligned}$$

luego

$$P[X=1 / Y=1] = \frac{\frac{1-\alpha}{6}}{\frac{1-\alpha}{6} + \frac{\beta}{6} + \frac{\gamma}{4}} = 2\frac{-1+\alpha}{-2+2\alpha-2\beta-3\gamma}.$$

3.9. Más sobre el Teorema de Bayes

La importancia del Teorema de Bayes en Estadística va mucho más allá de su aplicación como fórmula que facilita probabilidades condicionadas. La filosofía que subyace en él ha dado lugar a toda una forma de entender la Estadística, llamada por ello *Estadística Bayesiana*. Vamos a tratar de explicar los fundamentos de esta manera de entender el teorema.

Supongamos que hay un suceso A sobre el que tenemos un serio desconocimiento acerca de si se da o no se da. Tanto es así que tenemos que determinar la probabilidad de dicho suceso, $P[A]$. Es importante entender que nosotros somos conscientes de que A ha ocurrido o no ha ocurrido: el problema es precisamente que no sabemos qué ha pasado. Decimos que es importante porque $P[A]$ no representa la *probabilidad de que A ocurra*, sino nuestro grado de creencia en que ha ocurrido.

Es posible que no tengamos, en principio, datos para conocer de forma exacta cuál es la probabilidad de A. Aún así, podríamos atrevernos, *como expertos en el tema*, a dar una estimación de dicha probabilidad, $P[A]$. A esta probabilidad inicial que damos la vamos a llamar **probabilidad a priori.**

Ahora bien, hemos dado una probabilidad a priori $P[A]$ sin ninguna información sobre A. Supongamos ahora

que tenemos nueva información que nos dará pistas acerca de si A ha ocurrido o no, y que dicha información está recogida en un suceso que llamaremos B_1. En ese caso, podríamos y deberíamos *actualizar la probabilidad* de A basándonos en esta nueva información, proporcionando una nueva probabilidad de A que tenga en cuenta B_1, es decir, $P[A|_{B_1}]$, que llamaremos **probabilidad a posteriori**.

En esa *actualización de la probabilidad* es donde entra el Teorema de Bayes, ya que nos dice que

$$P[A|_{B_1}] = \frac{P[B_1|_A]P[A]}{P[B_1|_A]P[A] + P[B_1|_{\bar{A}}]P[\bar{A}]}.$$

Obsérvese que la probabilidad a posteriori es proporcional a la probabilidad a priori.

Finalmente, es muy importante ver que podemos extender esta forma de trabajar aplicando el teorema de una forma recursiva. Después de conocer B_1, nuestra nueva probabilidad para A es $P[A|_{B_1}]$. Abusando de la notación, podemos decir que esa es nuestra nueva probabilidad a priori y si, por ejemplo, tenemos más información sobre A, dada por otro suceso B_2, **información independiente de** B_1, la nueva probabilidad a posteriori sería

$$\begin{aligned} P[A|_{B_1\cap B_2}] &= \frac{P[B_2|_{A\cap B_1}]P[A|_{B_1}]}{P[B_2|_{A\cap B_1}]P[A|_{B_1}] + P[B_2|_{\bar{A}\cap B_1}]P[\bar{A}|_{B_1}]} \\ &= \frac{P[B_2|_A]P[A|_{B_1}]}{P[B_2|_A]P[A|_{B_1}] + P[B_2|_{\bar{A}}]P[\bar{A}|_{B_1}]}. \end{aligned}$$

Es muy importante observar que en este cociente $P[A|_{B_1}]$ ocupa el lugar que antes ocupaba la probabilidad a priori. Además, esta segunda probabilidad a posteriori podría considerarse como la nueva probabilidad a priori para una nueva aplicación del teorema basada en el conocimiento de nueva información dada por un suceso B_3. Este proceso de actualización de las probabilidades a priori basada en la información disponible puede realizarse cuantas veces sea necesario.

Vamos a ilustrar esto en un par de ejemplos.

3.9.1. Ejemplo del juez

Supongamos que un juez debe decidir si un sospechoso es inocente o culpable. Él sabe que debe ser cuidadoso y garantista con los derechos del acusado, pero también por su experiencia parte de una creencia en que el sospechoso puede ser culpable que, en cualquier caso, estima por debajo de lo que realmente cree para, insisto, ser garantista con los derechos del acusado. Pongamos que estima esta probabilidad en un 10 %.

Ahora empieza a examinar las pruebas. La primera de ellas es una prueba de ADN en la que el acusado dio positivo: encontraron material genético en el arma del crimen que, según la prueba, es suyo. Esa prueba de ADN da positivo en el 99.5 % de las veces en que se comparan dos ADN's idénticos, pero también da positivo (erróneamente) en el 0.005 % de las veces en que se aplica a dos ADN's distintos. Teniendo en cuenta esta información, el juez aplica por primera vez el teorema de Bayes con los siguientes datos:

- $P[culpable] = 0.1$, que es la probabilidad a priori que el juez considera.
- La probabilidad de que la prueba de ADN de positivo si el acusado es culpable es

$$P[ADN+|_{culpable}] = 0.995.$$

- La probabilidad de que la prueba de ADN de positivo si el acusado es inocente es

$$P\left[ADN+ \left|_{inocente}\right.\right] = 0.00005.$$

Ahora ya puede actualizar su grado de creencia en la culpabilidad del sospechoso:

$$\begin{aligned} P\left[culpable \left|_{ADN+}\right.\right] &= \frac{P\left[ADN+ \left|_{culpable}\right.\right] \times P\left[culpable\right]}{P\left[ADN+ \left|_{culpable}\right.\right] \times P\left[culpable\right] + P\left[ADN+ \left|_{inocente}\right.\right] \times P\left[inocente\right]} \\ &= \frac{0.995 \times 0.1}{0.995 \times 0.1 + 0.00005 \times 0.9} = 0.999548 \end{aligned}$$

Es decir, ahora piensa que el sospechoso es culpable con un 99.9548 % de certeza. Fijémonos en que nuestra probabilidad a priori aparece en los términos 0.1 en el numerador y 0.1 y 0.9 en el denominador. Esa, 0.1, era la probabilidad que teníamos **antes de la prueba** de que fuera culpable (y 0.9 de que fuera inocente); **después de la prueba** esa probabilidad es 0.999548 de que sea culpable (y 0.000452 de que sea inocente).

Sin embargo, el sospechoso insiste en su inocencia, y propone someterse a una prueba de un detector de mentiras. Los expertos saben que un culpable es capaz de engañar a esta máquina en el 10 % de las veces, y que la máquina dirá el 1 % de las veces que un inocente miente. Nuestro sospechoso se somete a la máquina y ésta dice que es inocente. ¿Cuál será ahora la probabilidad que el juez asigna a la culpabilidad del sospechoso? Teniendo en cuenta que:

- $P\left[maquina- \left|_{culpable}\right.\right] = 0.1,$
- $P\left[maquina+ \left|_{inocente}\right.\right] = 0.01,$

debe aplicar de nuevo el Teorema de Bayes, considerando ahora que la probabilidad a priori de que sea culpable es 99.9548 %:

$$\begin{aligned} P\left[culpable \left|_{maquina-}\right.\right] &= \frac{P\left[maquina- \left|_{culpable}\right.\right] \times P\left[culpable\right]}{P\left[maquina- \left|_{culpable}\right.\right] \times P\left[culpable\right] + P\left[maquina- \left|_{inocente}\right.\right] \times P\left[inocente\right]} \\ &= \frac{0.1 \times 0.999548}{0.1 \times 0.999548 + (1 - 0.01) \times (1 - 0.999548)} = 0.9955431. \end{aligned}$$

Es decir, aún con esa prueba negativa, el juez aún tiene un 99.55431 % de certidumbre de que el sospechoso es culpable. De nuevo, podemos resumir este paso diciendo que **antes de la segunda prueba** nuestra probabilidad de que fuera culpable era de 0.999548 (que aparece en la fórmula ocupando la posición de la probabilidad a priori), mientras que **después de la segunda prueba** esa probabilidad es 0.9955431.

El proceso puede verse resumido en el Cuadro 3.3.

3.9.2. Ejemplo de la máquina de detección de fallos

En un proceso industrial de producción en serie de capós de coche, existe una máquina encargada de detectar desperfectos que desechen una pieza de capó. Esa máquina está calibrada para detectar una pieza defectuosa con un 90 % de acierto, pero también detecta como defectuosas el 5 % de las piezas no defectuosas. El encargado de calidad estima, por estudios previos, que el porcentaje general de piezas defectuosas es del 5 %.

Este encargado, consciente de que la máquina puede dar por buenas piezas que son defectuosas, decide actuar de la siguiente forma: una pieza que sea detectada como no defectuosa pasará otras dos veces por la misma máquina detectora y sólo será declarada no defectuosa cuando en ninguna de esas tres pruebas, de defectuosa.

	$P\left[Culpable\right]$	
	Antes de la prueba	**Después de la prueba**
1ª prueba: $ADN+$	0.1	$\frac{P[ADN+\vert_{culpable}]\times 0.1}{P[ADN+\vert_{culpable}]\times 0.1+P[ADN+\vert_{inocente}]\times(1-0.1)}=0.999548$
2ª prueba: $maquina-$	0.999548	$\frac{P[maquina-\vert_{culpable}]\times 0.999548}{P[maquina-\vert_{culpable}]\times 0.999548+P[maquina-\vert_{inocente}]\times(1-0.999548)}=0.9955431$

Cuadro 3.3: Esquema del proceso iterativo del teorema de Bayes en el ejemplo del juez. La probabilidad *a priori* (antes de cada prueba) es la que se utiliza en la fórmula para obtener la probabilidad *a posteriori* (despúés de cada prueba). La probabilidad *a posteriori* (después) de una prueba es la probabilidad *a priori* (antes) de la siguiente prueba.

Supongamos que una pieza pasa las tres veces y da no defectuosa: ¿cuál es la probabilidad de que realmente sea no defectuosa?

Vamos a empezar notando adecuadamente los sucesos. Notaremos D al suceso ser defectuosa y por $+$ a dar positivo como defectuosa en la prueba de la máquina. Sabemos que:

- $P\left[D\right]=0.05$, que es la probabilidad a priori;
- $P\left[+\,|_{D}\right]=0.9$ y
- $P\left[+\,|_{\bar{D}}\right]=0.05$.

La probabilidad a priori de que una pieza sea no defectuosa es de 0.95, pero si es detectada como defectuosa una primera vez, dicha probabilidad pasa a ser

$$\begin{aligned}P\left[\bar{D}\,|_{\bar{+}}\right]&=\frac{P\left[\bar{+}\,|_{\bar{D}}\right]P\left[\bar{D}\right]}{P\left[\bar{+}\,|_{\bar{D}}\right]P\left[\bar{D}\right]+P\left[\bar{+}\,|_{D}\right]P\left[D\right]}\\&=\frac{0.95\times 0.95}{0.95\times 0.95+0.1\times 0.05}=0.9944904.\end{aligned}$$

Esa probabilidad pasa a ser la probabilidad a priori para la segunda vez que da no defectuosa. Por tanto, la probabilidad de que sea no defectuosa si da negativo por segunda vez es

$$\begin{aligned}P\left[\bar{D}\,|_{\bar{+}\bar{+}}\right]&=\frac{P\left[\bar{+}\,|_{\bar{D}}\right]0.9944904}{P\left[\bar{+}\,|_{\bar{D}}\right]0.9944904+P\left[\bar{+}\,|_{D}\right](1-0.9944904)}\\&=\frac{0.95\times 0.9944904}{0.95\times 0.9944904+0.1\times(1-0.9944904)}=0.9994172.\end{aligned}$$

Finalmente, la probabilidad de que sea no defectuosa si da negativo por tercera vez es

$$\begin{aligned}P\left[\bar{D}\,|_{\bar{+}\bar{+}\bar{+}}\right]&=\frac{P\left[\bar{+}\,|_{\bar{D}}\right]0.9994172}{P\left[\bar{+}\,|_{\bar{D}}\right]0.9994172+P\left[\bar{+}\,|_{D}\right](1-0.9994172)}\\&=\frac{0.95\times 0.9994172}{0.95\times 0.9994172+0.1\times(1-0.9994172)}=0.9999386.\end{aligned}$$

Como podemos ver, si una pieza da no defectuosa tres veces, la probabilidad de que sea realmente no defectuosa es altísima, del orden del 99.99 %, así que el método ideado por el responsable de calidad parece consistente.

	$P\left[\bar{D}\right]$	
	Antes de la prueba	**Después de la prueba**
1ª prueba: $\bar{+}$	0.95	$\frac{P[\bar{+}\|_{\bar{D}}]0.95}{P[\bar{+}\|_{\bar{D}}]0.95+P[\bar{+}\|_{D}](1-0.95)} = 0.9944904$
2ª prueba: $\bar{+}$	0.9944904	$\frac{P[\bar{+}\|_{\bar{D}}]0.9944904}{P[\bar{+}\|_{\bar{D}}]0.9944904+P[\bar{+}\|_{D}](1-0.9944904)} = 0.9994172$
3ª prueba: $\bar{+}$	0.9994172	$\frac{P[\bar{+}\|_{\bar{D}}]0.9994172}{P[\bar{+}\|_{\bar{D}}]0.9994172+P[\bar{+}\|_{D}](1-0.9994172)} = 0.9999386$

Cuadro 3.4: Esquema del proceso iterativo del teorema de Bayes en el ejemplo de la máquina de detección de fallos. La probabilidad *a priori* (antes de cada prueba) es la que se utiliza en la fórmula para obtener la probabilidad *a posteriori* (despúés de cada prueba). La probabilidad *a posteriori* (después) de una prueba es la probabilidad *a priori* (antes) de la siguiente prueba.

Capítulo 4

Variable aleatoria. Modelos de distribuciones de probabilidad

> Mas a pesar de todo eso, aunque la mala suerte exista, muy pocos reporteros veteranos creen de verdad en ella. En la guerra, las cosas suelen discurrir más bien según la ley de las probabilidades: tanto va el cántaro a la fuente que al final hace bang.
>
> Arturo Pérez Reverte, en *Territorio Comanche*

Resumen. En este capítulo continuamos con el estudio de la probabilidad, utilizando el concepto de variable aleatoria para referirnos a experimentos donde el resultado queda caracterizado por un valor numérico. Se presentan algunos de los modelos más habituales de asignación de probabilidades y sus propiedades más relevantes.

Palabras clave: variable aleatoria, variable discreta, función masa de probabilidad, variable continua, función de densidad de probabilidad, función de distribución, media, varianza, distribución binomial, distribución de Poisson, distribución geométrica, distribución uniforme, distribución exponencial, distribución Gamma, distribución normal.

4.1. Introducción

En el tema anterior hemos visto que la Estadística se ocupa de experimentos aleatorios. En general, en Ciencia y Tecnología se suele analizar cualquier experimento mediante una o varias medidas del mismo. Por ejemplo, se analiza un objeto según su peso, su volumen, su densidad, su contenido de agua...; o se analiza el tráfico de Internet según el número de conexiones a un servidor, el volumen total de tráfico generado, la velocidad...

En estos sencillos ejemplos observamos que se ha descrito un fenómeno físico, como puede ser un objeto o el estado de una red de comunicaciones en un momento dado, mediante uno o varios números o variables. Cuando ese fenómeno es de tipo aleatorio, vamos a llamar a esa asignación *variable aleatoria*.

Consideremos un experimento probabilístico con un espacio muestral Ω en el que se ha definido una función de probabilidad $P[\cdot]$.

Una **variable aleatoria** (a partir de ahora **v.a.**) es un número real asociado al resultado de un experimento aleatorio. Se trata, por tanto, de una función real con dominio en el espacio muestral, $X : \Omega \to \mathbb{R}$.

Podemos pensar en una v.a. como en una variable asociada a una población conceptual, ya que sólo podrá observarse cuando se tomen muestras suyas.

En la notación que vamos a utilizar representaremos las variables aleatorias como funciones siempre en mayúsculas, y a sus valores concretos siempre en minúscula. Es decir, si queremos referirnos a una v.a. antes de observar su valor, podemos notarla como X, por ejemplo; pero una vez que se observa el valor de dicha variable (ya no es, por tanto, algo aleatorio), debemos notar a ese valor en minúscula, por ejemplo, como x.

Por ejemplo, podemos decir que la variable aleatoria X que corresponde a la puntuación obtenida al lanzar el dado puede tomar los valores $x = 1, 2, 3, 4, 5, 6$. Podremos preguntarnos por la probabilidad de que X tome el valor $x = 4$ o de que $X \leq 6$. Si lanzamos el dado y observamos que ha salido un 6, diremos que $x = 6$.

No olvidemos que el objeto de la Estadística con respecto a la observación de fenómenos aleatorios es medir la certidumbre o la incertidumbre asociada a sus posibles resultados. Al describir estos resultados mediante variables aleatorias, lo que tenemos son resultados numéricos sujetos a incertidumbre. El objetivo ahora es cuantificar la probabilidad de esos resultados numéricos de alguna forma.

4.2. Variable aleatoria discreta

4.2.1. Definición

Se dice que una v.a. es **discreta** si el conjunto de todos los valores que puede tomar es un conjunto, a lo sumo, numerable (discreto).

Ejemplo. Son variables discretas:

- El número de accidentes laborales en una empresa al año.
- El número de errores en un mensaje transmitido.
- El número de piezas defectuosas producidas a lo largo de un día en una cadena de producción.
- El número de días de baja de un trabajador al mes.

4.2.2. Función masa de probabilidad

Dada una v.a. discreta, X, se define su **función masa de probabilidad** como

$$f(x) = P[X = x],$$

para cada $x \in \mathbb{R}$.

Nota. Obsérvese que una función masa de una v.a. discreta está definida en todos los puntos de la recta real, pero sólo valdrá distinto de cero en un conjunto, a lo sumo, numerable, que corresponde con los únicos valores que pueden darse de la variable.

Sea X una v.a. discreta y $f(x)$ su función masa. Entonces:

1. $f(x) \geq 0$ para todo $x \in \mathbb{R}$.
2. $\sum_{x\in\mathbb{R}} f(x) = 1$.
3. En general, para cualquier conjunto B,

$$P[X \in B] = \sum_{x_i \in B} f(x_i),$$

donde x_i son valores posibles de X.

4.2.3. Función masa de probabilidad empírica

En la práctica nadie conoce la auténtica función masa de una variable discreta, pero podemos aproximarla mediante la *función masa de probabilidad empírica* asociada a una muestra de resultados.

Si tenemos una colección de posibles resultados de la variable X, $x_1, ..., x_N$, esta función asigna al valor x la frecuencia con la que dicho valor se da en la muestra, es decir,

$$f_{emp}(x) = \frac{\textit{número de valores}\, x_i \textit{ iguales a } x}{N}.$$

Si el tamaño, N, de la muestra es grande, esta función tiende a la auténtica, es decir, para cada $x \in \mathbf{R}$.

$$\lim_{N\to\infty} f_{emp}(x) = f(x).$$

Ejemplo. En la Figura 4.1 aparece la función masa empírica correspondiente al lanzamiento de un dado 600 veces. Esta función empírica aparece representada en barras verticales, mientras que la función masa *teórica*, $f(x) = \frac{1}{6}$, para $x = 1, 2, 3, 4, 5, 6$ aparece representada como una línea horizontal. Puede apreciarse cómo proporcionan probabilidades teóricas y empíricas bastante parecidas. No obstante, ¿deberíamos concluir a la luz de estos 600 datos que el dado no está cargado?

4.2.4. Media y varianza de una variable aleatoria discreta

Dada una v.a. discreta, X, con función masa de probabilidad $f(x)$, se define su media o esperanza matemática como

$$EX = \sum_x x \times f(x).$$

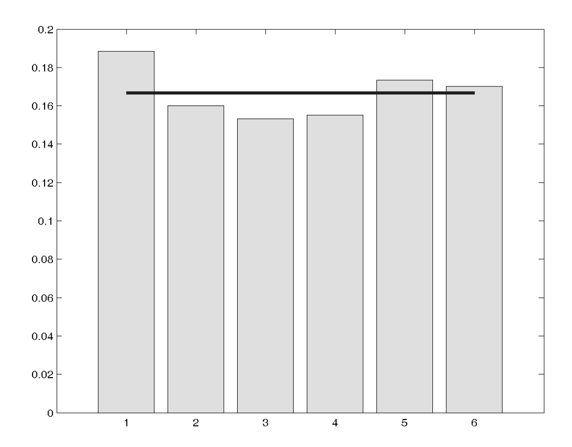


Figura 4.1: Función masa empírica de una muestra de 600 lanzamientos de un dado.

Como en el caso de la media muestral de unos datos, la media de una v.a. se interpreta como el centro de gravedad de los valores que puede tomar la variable, con la diferencia que en una media muestral, el *peso* de cada valor lo da la frecuencia de dicho valor en los datos y aquí el *peso* lo determina la probabilidad, dada por la función masa.

Dada una v.a. discreta, X, con función masa de probabilidad $f(x)$, se define su varianza como

$$VarX = \sum_x (x - EX)^2 \times f(x).$$

La forma más cómoda de calcular en la práctica la varianza es desarrollando previamente el cuadrado que aparece en su definición, ya que

$$\begin{aligned} VarX =& \sum_x (x - EX)^2 \times f(x) = \sum_x (x^2 - 2xEX + EX^2) \times f(x) \\ =& \sum_x x^2 \times f(x) - 2EX \times \sum_x x \times f(x) + EX^2 \times \sum_x f(x) \\ =& E[X^2] - 2EX^2 + EX^2 = E[X^2] - EX^2. \end{aligned}$$

Al igual que ocurre con la varianza muestral es conveniente definir la desviación típica de una v.a., como $\sigma = \sqrt{VarX}$, que tiene las mismas unidades que la media y que se puede interpretar como una media del grado de variación del conjunto de valores que puede tomar la v.a. respecto del valor de la media.

4.3. Modelos de distribuciones de probabilidad para variables discretas

Según lo que hemos visto hasta ahora, la forma en que se asigna probabilidad a los resultados de una variable aleatoria discreta viene dada por la función masa de probabilidad. A esta manera de determinar la

probabilidad asociada a los resultados de la variable la vamos a llamar a partir de ahora **distribución de probabilidad** de una v.a. Démonos cuenta que, como acabamos de comentar, para determinar la distribución de probabilidad de una v.a. sólo tenemos que dar su función función masa de probabilidad.

Sin embargo, debemos tener en cuenta que en la vida real nadie conoce cuál es la auténtica distribución de probabilidad de una v.a., porque nadie sabe a priori cuál es la función masa de dicha variable. Todo lo más, podemos calcular la función masa empírica a partir de los datos de una muestra. Aún así, llegará el momento de *pasar al límite*, es decir, de inducir una fórmula teórica que corresponda a la distribución de probabilidad que proponemos y que se parezca a la distribución empírica de los datos de la muestra.

Para ayudar a ese *paso al límite*, en Estadística se estudian **modelos teóricos de distribuciones de probabilidad**. Se trata de fórmulas teóricas de funciones masa que pueden resultar adecuadas para determinadas variables aleatorias.

Hay una metáfora que puede ayudar a entender cómo se asigna una distribución de probabilidad y sobre la que abundaremos en lo sucesivo: ¿qué ocurre cuando queremos comprar unos pantalones? En general acudimos a una tienda de moda y:

1. De entre una serie de modelos, elegimos el modelo que creemos que mejor nos va.
2. Buscamos la talla que hace que mejor se ajuste a nosotros, según nuestras características.

Pues bien, en el caso de las v.a.

- *nuestras características* son las posibles observaciones que tenemos sobre la v.a. que, por ejemplo, pueden determinar una distribución empírica asociada a una muestra;
- *los modelos* de la tienda, entre los que elegimos el que más nos gusta, son los modelos teóricos que vamos a empezar a estudiar a continuación;
- y *la talla* que hace que los pantalones se ajusten a nosotros adecuadamente son los parámetros de los modelos teóricos.

En lo que resta de este capítulo vamos a describir algunos de los modelos teóricos de probabilidad más habituales en el ámbito de las Ingenierías, comenzando por el caso de v.a. discretas.

4.3.1. Distribución binomial

Sea X una v.a. discreta que toma los valores $x = 0, 1, ..., n$, donde n es un número natural conocido. Se dice que X sigue una **distribución binomial de parámetros** n y p (y se nota $X \to B(n, p)$) si su función masa es

$$\begin{aligned} f(x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, ..., n. \end{aligned}$$

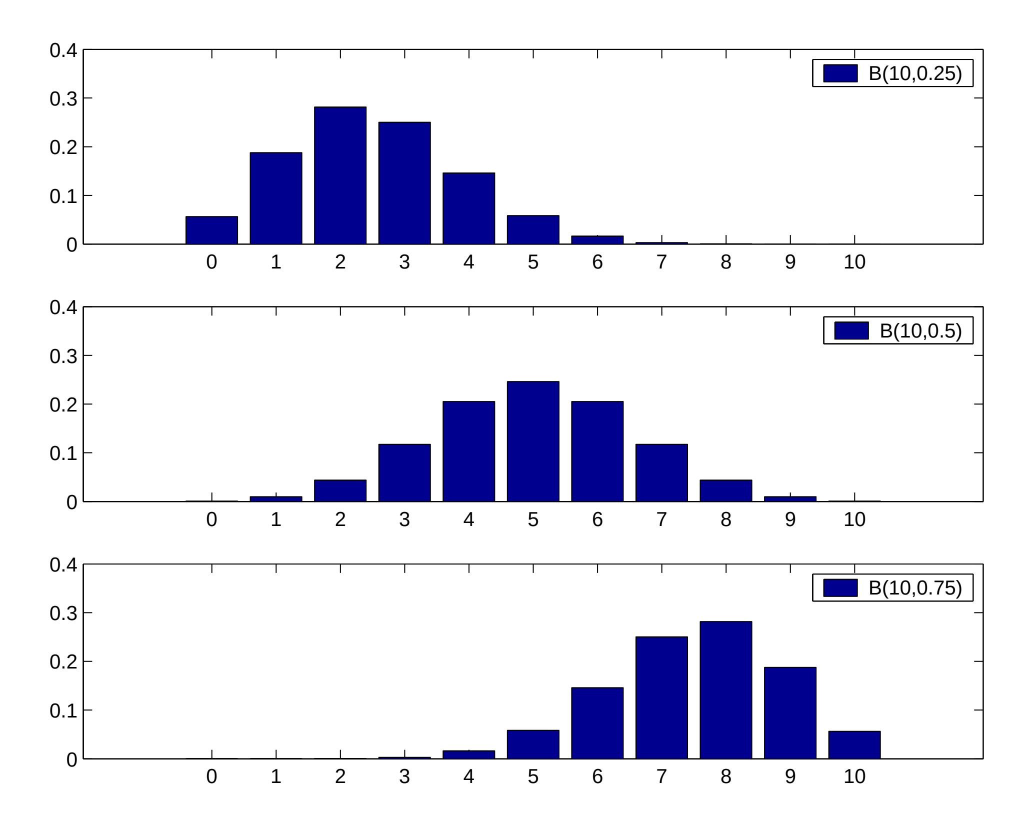


Figura 4.2: Funciones masa de distribuciones binomiales.

Sea $X \to B(n,p)$. Entonces

$$EX = np$$
$$VarX = np(1-p).$$

Caracterización de la distribución binomial. Supongamos que un determinado experimento aleatorio se repite n veces de forma independiente y que en ese experimento hay un suceso que denominamos *éxito,* que ocurre con probabilidad constante p. En ese caso, la variable aleatoria X que mide el número de éxitos sigue una $B(n,p)$.

En esta caracterización es importante observar que las dos hipótesis fundamentales de esta distribución son:

- los experimentos se repiten de forma **independiente** y
- la probabilidad de éxito es **constante**.

En la medida en que estas dos hipótesis no sean válidas, la distribución binomial no será adecuada para la variable que cuenta el número de éxitos.

Un ejemplo particular de distribución binomial lo constituye la denominada **distribución de Bernouilli**. Se trata de una distribución $B(1,p)$, con función masa

$$f(x) = \begin{cases} 1-p \text{ si } x=0 \\ p \text{ si } x=1 \end{cases}.$$

x	0	1	2	3	4
$P[X=x]$	$\binom{4}{0}0.2^0 0.8^4$ $=0.41$	$\binom{4}{1}0.2^1 0.8^3$ $=0.41$	$\binom{4}{2}0.2^2 0.8^2$ $=0.15$	$\binom{4}{3}0.2^3 0.8^1$ $=0.03$	$\binom{4}{4}0.2^4 0.8^0$ $=0.00$

Cuadro 4.1: Función masa de una $B(4, 0.2)$

Ejemplo. Consideremos como v.a. el número de días a la semana que un joven de hoy consume alcohol. ¿Podríamos pensar que se trata de una v.a. con distribución $B(7, p)$, donde $p = \frac{\textit{número medio de días de consumo}}{7}$? Probablemente no, porque

1. Puede darse el *efecto resaca*, es decir, si se consume mucho un día, huir del alcohol al día siguiente; o el efecto inverso *un clavo quita otro clavo*; o ...; en definitiva, circunstancias que rompan la hipótesis de independencia en el consumo en días distintos.

2. Está claro que la probabilidad de consumir un martes no es, en general, la misma que un sábado. Tampoco todos los jóvenes tienen la misma probabilidad de consumir alcohol un día cualquiera.

Ejemplo. Un ingeniero se ve obligado a transmitir dígitos binarios a través de un sistema de comunicaciones bastante imperfecto. Por estudios previos, estima que la probabilidad de que un dígito se transmita incorrectamente es del 20 %. El ingeniero envía un mensaje de 4 dígitos y se pregunta cuántos se recibirán incorrectamente.

Desde el punto de vista estadístico nosotros no podemos responder a esa pregunta. En realidad, nadie puede responder a esa pregunta con certeza, porque existe incertidumbre latente en ella: el azar determinará cuántos dígitos se cruzan. Lo que sí podemos hacer es facilitarle el grado de certeza, es decir, la probabilidad, de cada uno de los posibles resultados.

Concretamente, si analizamos la variable X: *número de dígitos que se reciben incorrectamente*, teniendo en cuenta que el ensayo de cada envío de cada dígito se hará de forma independiente y que nos ha dicho que la probabilidad de que un dígito se reciba incorrectamente es 0.2, podemos afirmar que un modelo de probabilidad adecuado para dicha variable es una distribución $B(4, 0.2)$. Esta distribución nos permite calcular la probabilidad de que se crucen 0, 1, 2, 3 o 4 de los dígitos. Lo esquematizamos en la tabla adjunta. Vistos los resultados, debemos decirle al ingeniero que es hartamente improbable que le fallen los 4 dígitos, pero que tiene una probabilidad (ver Cuadro 4.1) de

$$0.41 + 0.15 + 0.03 + 0.00 = 0.59$$

de que le falle el envío de al menos uno de ellos.

4.3.2. Distribución de Poisson

Sea X una v.a. discreta, que puede tomar los valores $x = 0, 1, 2, ...$ Se dice que X sigue una **distribución de Poisson de parámetro** λ (y se nota $X \to P(\lambda)$) si su función masa es

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \ \ x = 0, 1, 2, ...$$

Sea $X \to P(\lambda)$. Entonces

$$EX = \lambda$$
$$VarX = \lambda.$$

Caracterización de la distribución de Poisson. Consideremos el número de éxitos en un periodo de tiempo donde los éxitos acontecen a razón de λ veces por unidad de tiempo (en promedio) y de forma independiente. En ese caso

$$X : \textit{número de ocurrencias del suceso por unidad de tiempo}$$

es una variable de **Poisson de parámetro** λ, y se nota $X \to P(\lambda)$.

En esta caracterización, las hipótesis fundamentales ahora son:

- la **independencia** de las realizaciones y
- el promedio **constante** de ocurrencias por unidad de tiempo.

Ejemplo. La distribución de Poisson suele utilizarse como modelo para el número de accidentes ocurridos en los individuos de una población a lo largo de un periodo de tiempo. Lo que mucha gente no termina de asumir es que hacer esa suposición equivale a decir que todos esos individuos tienen el mismo riesgo de tener un accidente y que el hecho de que un individuo tenga un accidente no modifica para nada la probabilidad de sufrir un nuevo accidente. Es evidente que en muchas situaciones de la vida real eso no es cierto, así que el modelo no será adecuado en ellas.

Ejemplo. Otra aplicación muy común de la distribución de Poisson es al número de partículas por unidad de volumen en un fluido cuando una disolución está realmente bien disuelta. En caso de que los datos indiquen que la distribución de Poisson no es adecuada, podríamos de hecho inferir que la disolución no está bien disuelta.

Ejemplo. En el contexto de las redes de telecomunicaciones, el uso más común de la distribución de Poisson es en el ámbito del número de solicitudes de servicio a un servidor. Por ejemplo, se suele considerar que el nº de llamadas a una centralita o el nº de conexiones a un servidor sigue una distribución de Poisson.

Sin embargo, hay que decir que aunque este uso de la distribución de Poisson es muy común, es evidente que la hipótesis de que el promedio λ debe ser constante, no se da en estas aplicaciones, ya que uno de los fenómenos más conocidos en telecomunicaciones es el de *la hora cargada*: no es el mismo promedio de llamadas el que se produce a las 12 del mediodía que a las 3 de la mañana. Lo que se suele hacer es aplicar uno de los principios más importantes aunque menos escritos de la ingeniería, la ley de Murphy (*si algo puede ir mal, prepárate para ello, porque en algun momento irá mal*): así, las redes de telecomunicaciones suelen dimensionarse para ser capaces de funcionar en el peor de los escenarios posibles, es decir, cuando el promedio de solicitudes es el que se da en la hora cargada.

Aproximación de la binomial. Ley de eventos raros. Supongamos que, como en la caracterización de la distribución binomial, un determinado experimento aleatorio se repite n veces de forma independiente y que en ese experimento hay un suceso que denominamos *éxito,* que ocurre con probabilidad constante p. Adicionalmente, supongamos que el experimento se repite un gran número de veces, es decir, n es grande y que el éxito es un suceso raro, es decir, p es pequeño, siendo el promedio de ocurrencias, $\mu = np$. En ese caso, la variable aleatoria X que mide el número de éxitos sigue (aproximadamente) una $P(\mu)$.

En esta segunda caracterización se suele considerar aceptable la aproximación si $n > 20$ y $p < 0.05$. Si $n > 100$, la aproximación es generalmente excelente siempre y cuando $np < 10$. Hay que tener en cuenta que para esos valores de los parámetros, la distribución binomial tendría bastantes problemas para ser computada, ya que se exigiría, entre otros cálculos, el cálculo de $n!$ para un valor de n alto, por lo que la aproximación es muy útil.

Ejemplo. Supongamos que un fabricante de maquinaria pesada tiene instalados en el campo 3840 generadores de gran tamaño. Si la probabilidad de que cualquiera de ellos falle durante el año en curso es de $\frac{1}{1200}$, determinemos la probabilidad de que

a. 4 generadores fallen durante el año en curso,

b. Más 1 de un generador falle durante el año en curso.

El promedio de motores que fallan en el año es $\lambda = np = (3840)(1/1200) = 3.2$.

Sea X la variable que define el número de motores que pueden fallar en el año, con valores $x = 0, 1, 2, 3, ..., 3840$.

En principio, $X \to B(3840, 1/1200)$, pero dado que n es muy grande y p muy pequeño, podemos considerar que $X \to P(3.2)$. Por tanto,

$$P[X = 4] = \frac{e^{-3.2}3.2^4}{4!} = 0.178\,09$$

Por su parte,

$$P[X > 1] = 1 - P[X = 0, 1] = 1 - \frac{e^{-3.2}3.2^0}{0!} - \frac{e^{-3.2}3.2^1}{1!} = 0.828\,80$$

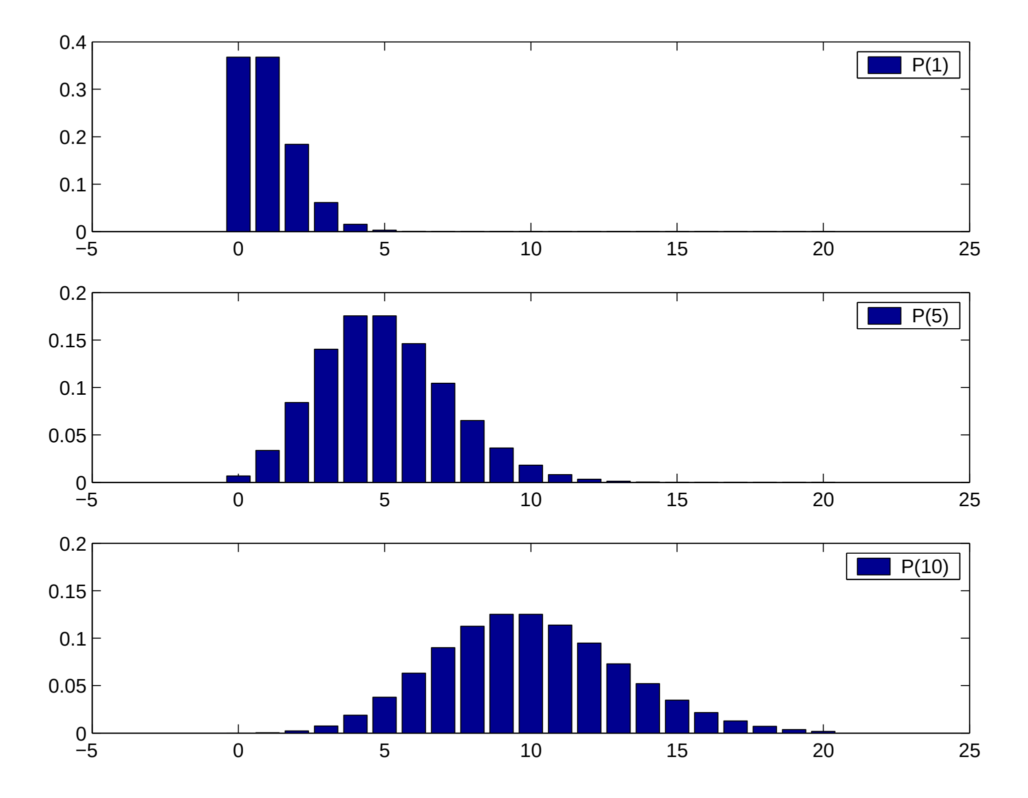


Figura 4.3: Funciones masa de distribuciones de Poisson.

4.3.3. Distribución geométrica

Sea X una v.a. discreta que puede tomar los valores $x = 0, 1, 2, ...$ Se dice que sigue una **distribución geométrica** de parámetro p (y se nota $X \to Geo\,(p)$), con $0 < p < 1$, si su función masa es

$$f\,(x) = p\,(1-p)^x\,,\ \text{para } x = 0, 1, 2, ...$$

Sea $X \to Geo\,(p)$. Entonces,

$$EX = \frac{1-p}{p}$$
$$VarX = \frac{1-p}{p^2}.$$

Caracterización de la distribución geométrica. Supongamos que un determinado experimento aleatorio se repite sucesivamente de forma independiente y que en ese experimento hay un suceso que denominamos *éxito,* que ocurre con probabilidad constante p. En ese caso, la variable aleatoria X que cuenta el número de fracasos hasta que ocurre el primer éxito sigue una $Geo\,(p)$.

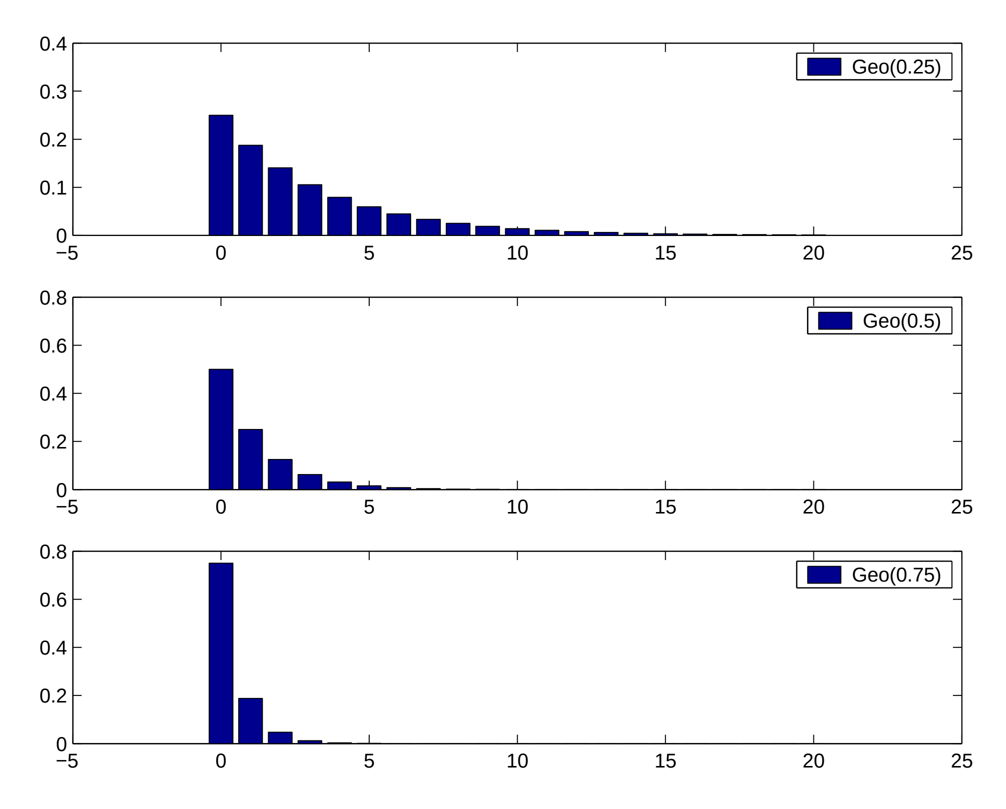


Figura 4.4: Funciones masa de distribuciones geométricas.

Ejemplo. Siguiendo con un ejemplo anterior, sobre el ingeniero que envía dígitos a través de un canal imperfecto, ahora se plantea cuántos dígitos se recibirán correctamente hasta que uno se cruce, sabiendo que la probabilidad de que uno cualquiera lo haga es de 0.2.

La variable de interés ahora es Y: *nº de dígitos que se reciben bien hasta el primero que se cruza*. Esta variable tiene como modelo de probabilidad una distribución $Geo(0.2)$. Gracias a este modelo, podemos decirle, por ejemplo, que la probabilidad de que envíe bien dos y que falle el tercero es de

$$P[Y = 2] = 0.2 \times 0.8^2 = 0.128.$$

4.3.4. Distribución binomial negativa

Sea una v.a. discreta que puede tomar los valores $x = 0, 1, 2, ...$ Se dice que X sigue una **distribución binomial negativa** de parámetros a y p (y se nota $X \to BN(a, p)$), con $a > 0$ y $0 < p < 1$, si su función masa es

$$f(x) = \frac{\Gamma(a + x)}{\Gamma(a)\Gamma(x + 1)} p^a (1 - p)^x \quad \text{para } x = 0, 1, 2, ...$$

donde $\Gamma(x) = \int_0^\infty s^{x-1} e^{-s} ds$ es la función gamma.

Obsérvese que la distribución geométrica es un caso particular de la binomial negativa, cuando $a = 1$.

Sea $X \to BN(a,p)$. Entonces

$$EX = a\frac{1-p}{p}$$
$$VarX = a\frac{1-p}{p^2}$$

Caracterización de la distribución binomial negativa. Sea un determinado experimento aleatorio que se repite sucesivamente de forma independiente y donde hay un suceso que denominamos *éxito,* que ocurre con probabilidad constante p. En ese caso, la variable aleatoria X que cuenta el número de fracasos hasta que ocurre el k-ésimo éxito sigue una $BN(k,p)$. En este caso, además, y dado que $\Gamma(r) = (r-1)!$ si r es un entero,

$$f(x) = \frac{(k+x-1)!}{(k-1)!x!}p^k(1-p)^x \quad \text{para } x = 0,1,2,...$$
$$= \binom{k+x-1}{k-1}p^k(1-p)^x \quad \text{para } x = 0,1,2,...$$

Caracterización de la distribución binomial negativa. Sean $X_1,...,X_n$ v.a. independientes[a] con distribución $Geo(p)$. En ese caso, $X = \sum_{i=1}^n X_i$ sigue una $BN(n,p)$. De nuevo obsérvese que el primer parámetro es un entero.

[a]Podemos quedarnos por ahora con la idea de que v.a. independientes son aquellas tales que el resultado de cualquiera de ellas no afecta al resto.

Ejemplo. Continuando con el ejemplo de la transmisión de dígitos a través de un sistema imperfecto, ¿cuántos dígitos se transmitirán correctamente hasta que dos lo hagan incorrectamente? De nuevo tenemos que asumir que no hay una respuesta para esto, pero sí podemos considerar un modelo de probabilidad para ello que nos ayude a tomar decisiones.

Sea Z: *nº de dígitos que se reciben bien hasta que dos se cruzan.* Esta v.a. sigue una distribución $BN(2,0.2)$. Gracias a este modelo, podemos decirle al ingeniero, por ejemplo, que la probabilidad de que se le crucen 2 dígitos con 10 o menos envíos es

$$P[Z \leq 8] = \sum_{z=0}^{8} P[Z = z] = \sum_{z=0}^{8} \frac{(2+z-1)!}{(2-1)!z!}0.2^2 0.8^z = 0.62$$

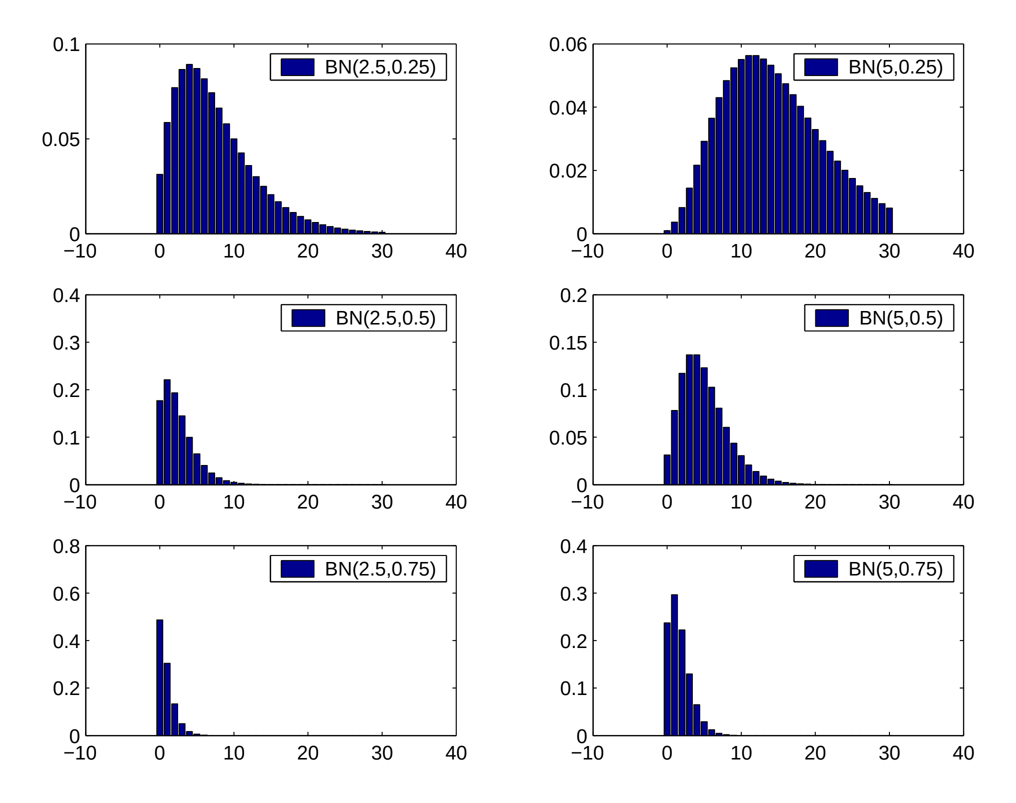


Figura 4.5: Funciones masa de distribuciones binomiales negativas.

4.4. Variable aleatoria continua

4.4.1. Definición

Una variable aleatoria es **continua** si el conjunto de valores que puede tomar sólo puede encerrarse en intervalos, formando, por tanto, un conjunto con un número infinito no numerable de elementos.

Ejemplo. Son variables aleatorias continuas:

- La tensión de fractura de una muestra de asfalto.
- El grosor de una lámina de aluminio.
- El pH de una muestra de lluvia.
- La duración de una llamada telefónica.

4.4.2. Histograma

Hay una diferencia fundamental entre las variables discretas y las continuas: en las discretas podemos, al menos, numerar los posibles valores y contar el número de veces que sale cada valor posible en una muestra. Sin embargo, por el carácter que tienen los intervalos de números reales, por muy grande que fuera la muestra

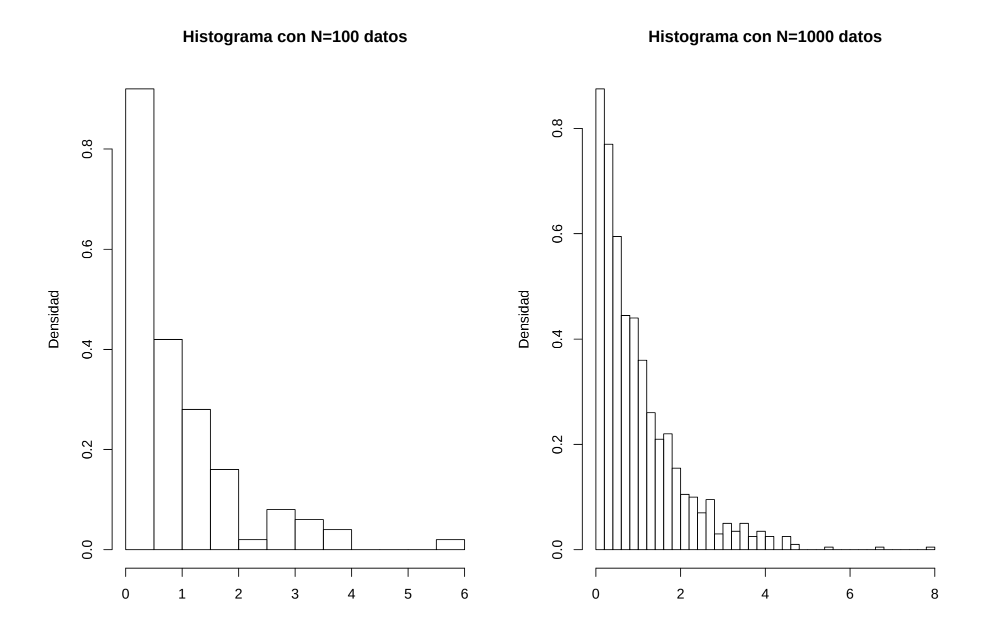


Figura 4.6: Histogramas.

que tomáramos de una variable continua, jamás tendríamos más de un valor de algunos puntos que puede tomar la variable[1].

Por esa razón, en una variable continua no podemos definir una función masa empírica, precisamente porque los valores de una variable continua no tienen masa de probabilidad.

Sin embargo, como sabemos, existe una representación análoga a la función masa empírica que permite aproximar las probabilidades de los valores de una variable continua: el histograma.

Vamos a considerar un sencillo ejemplo para ilustrar esta cuestión: mediante R simulamos dos muestras de una variable, una con $N = 100$ valores y otra con $N = 1000$. Histogramas asociados a estas muestras, con 10 y 31 intervalos, respectivamente, aparecen en la Figura 4.6. Teniendo en cuenta que el área de las barras representa la frecuencia relativa con que se dan los valores de los sucesivos intervalos en la muestra, en estos histogramas podemos ver que la variable toma mayoritariamente valores cercanos a cero; tanto más lejano al cero es un valor, menos probable parece ser. Este descenso de la probabilidad es además, muy acusado, casi exponencial.

Por otra parte, obsérvese que al pasar de 100 datos en la muestra a 1000 datos, el histograma esboza la forma de una función real de variable real. En general, cuanto mayor es N más se aproximan los histogramas a la forma de una función continua. Vamos a ir viendo cuál es la utilidad de esa función desde el punto de vista del Cálculo de Probabilidades.

Si en el histograma de la izquierda de la Figura 4.6 quisiéramos calcular la probabilidad en la muestra de alguno de los intervalos que definen el gráfico, la respuesta sería el área de la barra sobre dicho intervalo. Si quisiéramos la probabilidad en la muestra de varios intervalos, sumaríamos las áreas de las barras.

El problema es que para que las probabilidades en la muestra se parezcan a las verdaderas probabilidades es necesario que el tamaño de la muestra sea grande, cuanto mayor, mejor. En ese caso, tendríamos un

[1] Esto sucedería siempre que tomemos un número suficiente de decimales en cada valor.

histograma más parecido al de la derecha de la Figura 4.6. En él, de nuevo, si queremos, por ejemplo, calcular

$$P\left[a<X<b\right],$$

deberíamos sumar las áreas de las barras que forman el intervalo (a,b), si es que hay intervalos que forman, exactamente, el intervalo (a,b).

Pero si el tamaño de la muestra es lo suficientemente amplio para poder *pasar al límite* y encontrar una función real de variable real $f(x)$ que represente la línea que define el histograma, calcular una probabilidad del tipo $P[a<X<b]$ sumando las áreas de las barras de los intervalos infinitesimales que forman el intervalo (a,b) equivale a integrar dicha función en el intervalo (a,b), es decir,

$$P\left[a<X<b\right]=\int_a^b f(x)\,dx.$$

4.4.3. Función de densidad

Dada una v.a. continua, X, la **función de densidad de probabilidad** de X es aquella función $f(x)$ tal que para cualesquiera $a,\ b\in\mathbb{R}$ o $a,b=\pm\infty$,

$$P\left[a<X<b\right]=\int_a^b f(x)\,dx$$

Nota. Dado que a efectos del cálculo de integrales un punto no afecta al resultado de la integral, si $a,b\in\mathbf{R}$, podemos decir que

$$P\left[a<X<b\right]=\int_a^b f(x),$$
$$P\left[a\leq X<b\right]=\int_a^b f(x),$$
$$P\left[a<X\leq b\right]=\int_a^b f(x),$$
$$P\left[a\leq X\leq b\right]=\int_a^b f(x).$$

Este hecho pone de manifiesto que los valores concretos de una variable aleatoria continua no tienen masa de probabilidad, ya que

$$P\left[X=x_0\right]=\int_{x_0}^{x_0} f(x)\,dx=0,$$

pero sí tienen densidad de probabilidad, $f(x_0)$. Esta densidad de probabilidad representa la probabilidad de los intervalos infinitesimales de valores alrededor de x_0. Así, aunque $P[X=x_0]=0$, si $f(x_0)$ toma un valor alto, querrá decir que los valores alrededor de x_0 son muy probables.

Dada una v.a. continua, X con función de densidad $f(x)$:

1. $f(x) \geq 0$ para todo $x \in \mathbb{R}$.
2. $\int_{-\infty}^{\infty} f(x) = 1$.
3. En general, para cualquier conjunto de números reales, B,

$$P[X \in B] = \int_B f(x)\,dx.$$

4.4.4. Función de distribución

Se define la **función de distribución de probabilidad de una v.a. continua** X como

$$F(x) = P[X \leq x] = \int_{-\infty}^{x} f(t)\,dt.$$

Si X es una v.a. continua con función de densidad $f(x)$ y función de distribución $F(x)$, entonces

1. $\lim_{x\to-\infty} F(x) = 0$.
2. $\lim_{x\to\infty} F(x) = 1$.
3. F es creciente.
4. F es continua.
5. $f(x) = F'(x)$.

Ejemplo. Considérese una variable aleatoria continua, X, con función de densidad $f(x) = ce^{-a|x|}$. Vamos a calcular la constante c, la función de distribución y $P[X \geq 0]$.

En primer lugar,

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{\infty} f(x)\,dx \\ &= \int_{-\infty}^{0} c\exp(ax)\,dx + \int_{0}^{\infty} c\exp(-ax)\,dx = \frac{2c}{a}, \end{aligned}$$

luego es necesario que $c = \frac{a}{2}$.

Por otra parte,

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \begin{cases} \frac{1}{2}e^{ax} \text{ si } x < 0 \\ \frac{1}{2} + \frac{1-e^{-ax}}{2} \text{ si } x \geq 0 \end{cases}$$

Por último, $P[X \geq 0] = \int_0^{\infty} f(x)\,dx = \frac{1}{2}$.

La función de densidad y la de distribución, para $a = 1$, aparecen en la Figura 4.7.

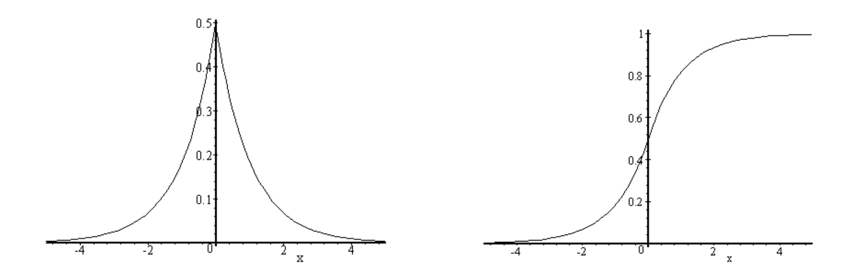


Figura 4.7: Función de densidad (izquierda) y de distribución (derecha).

Ejemplo. Consideremos una v.a. continua con función de distribución dada por

$$F(x) = \begin{cases} 0 \text{ si } x < 0 \\ x \text{ si } 0 \leq x < 1 \\ 1 \text{ si } x \geq 1 \end{cases}.$$

En ese caso, la función de densidad es

$$f(x) = F'(x) = \begin{cases} 1 \text{ si } 0 \leq x \leq 1 \\ 0 \text{ en otro caso} \end{cases}$$

Gráficamente, ambas funciones aparecen en la Figura 4.8. En esta variable, todos los puntos tienen la misma densidad de probabilidad, indicando que todos los intervalos de la misma longitud, dentro de $[0,1]$, tienen la misma probabilidad.

4.4.5. Función de distribución empírica

Al igual que ocurre con la función masa empírica con respecto a la función masa y al histograma con respecto a la función de densidad, la función de distribución, indistintamente de que se trate de una variable discreta o continua, también tiene una *versión muestral.*

Concretamente, si tenemos una variable aleatoria X y una muestra suya de tamaño N, $(x_1, ..., x_N)$, la **función de distribución empírica** se define como

$$S_N(x) = \frac{\textit{número de valores } \leq x}{N}.$$

Esta función se utiliza para aproximarse a la función de distribución, ya que para un gran número de valores,

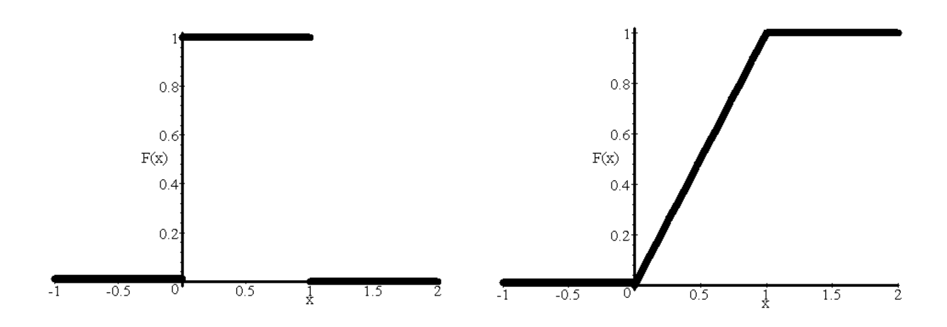


Figura 4.8: Función de densidad (izquierda) y de distribución (derecha).

la curva empírica se parecerá bastante a la función de distribución. Dicho de otra forma,

$$\lim_{N\to\infty} S_N(x) = F(x),$$

para cada x.

Ejemplo. En el ejemplo anterior se hablaba de una variable aleatoria continua cuya función de distribución es

$$F(x) = \begin{cases} 0 \text{ si } x < 0 \\ x \text{ si } x \in [0,1] \\ 1 \text{ si } x > 1 \end{cases}.$$

En la Figura 4.9 hemos representado dos funciones de distribución empíricas asociadas a sendas muestras de tamaño $N = 10$ (izquierda) y $N = 100$ (derecha).

Obsérvese que cuando aumenta el tamaño de la muestra (N), la función de distribución empírica se parece cada vez más a la función de distribución.

4.4.6. Media y varianza de una v.a. continua

Sea X una v.a. continua con función de densidad $f(x)$. Se define su media o esperanza matemática como

$$EX = \int_{-\infty}^{\infty} x \times f(x) dx.$$

La interpretación de la media de una v.a. continua es, de nuevo, la de un valor central alrededor del que se dan el conjunto de realizaciones de la v.a. Otra interpretación es la de **valor esperado**, en el sentido de que

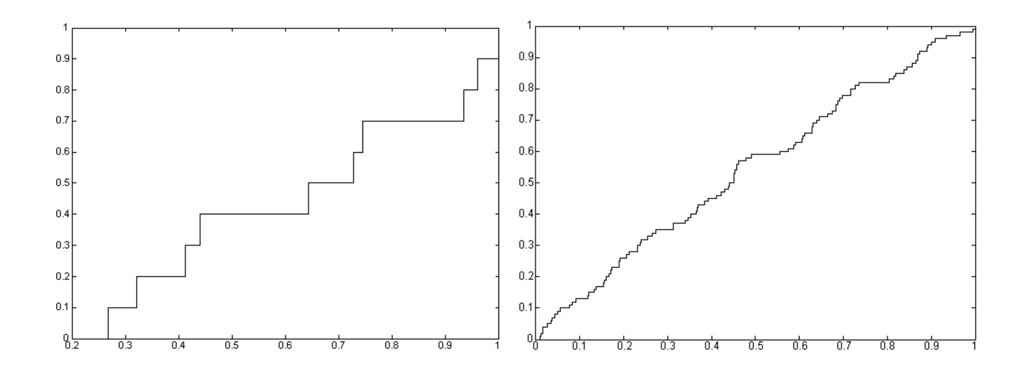


Figura 4.9: Funciones de distribución empíricas.

es el valor de la variable aleatoria en el que a priori se tienen más esperanzas.

Ejemplo. Sea una v.a. continua con función de densidad

$$f_X(x) = \begin{cases} \frac{1}{x_2 - x_1} \text{ si } x_1 \leq x \leq x_2 \\ 0 \text{ en otro caso} \end{cases}.$$

Calculemos su media:

$$\begin{aligned} EX &= \int_{x_1}^{x_2} x \cdot \frac{1}{x_2 - x_1} \cdot dx \\ &= \frac{1}{x_2 - x_1} \cdot \left[\frac{x^2}{2}\right]_{x_1}^{x_2} = \frac{1}{2} \cdot \frac{x_2^2 - x_1^2}{x_2 - x_1} \\ &= \frac{1}{2} \cdot \frac{(x_2 - x_1) \cdot (x_2 + x_1)}{x_2 - x_1} = \frac{1}{2}(x_1 + x_2), \end{aligned}$$

es decir, el punto medio del intervalo $[x_1, x_2]$.

Ejemplo. Sea una v.a. continua con función de densidad

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} \text{ si } x \geq 0 \\ 0 \text{ en otro caso} \end{cases}.$$

Calculemos su media:

$$EX = \int_0^\infty x \cdot \lambda \cdot e^{-\lambda x} \cdot dx$$

$$\begin{array}{c} u = x \\ dv = \lambda \cdot e^{-\lambda x} \cdot dx \\ = \end{array} \quad \left[-x \cdot e^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x} \cdot dx$$

$$= 0 + \left[-\frac{1}{\lambda} e^{-\lambda x}\right]_0^\infty = \frac{1}{\lambda}.$$

Vamos a introducir ahora el concepto de varianza de una v.a. continua, que de nuevo se interpreta como una medida de la concentración de los valores de la v.a. en torno a su media.

Sea una v.a. X. Se define su **varianza** como $Var[X] = E\left[(X - EX)^2\right]$.

Es decir, es la media de las desviaciones al cuadrado de los valores de la variable respecto de su media.

La raíz cuadrada de la varianza, $\sigma = \sqrt{Var[X]}$ se conoce como **desviación típica**.

Como en el caso de las v.a. discretas, existe un método más cómodo para el cálculo de cualquier varianza. En concreto,

$$\begin{aligned} Var[X] &= E\left[(X - EX)^2\right] = E\left[X^2 - 2X \cdot EX + (EX)^2\right] \\ &= E\left[X^2\right] - 2 \cdot EX \cdot EX + (EX)^2 = E\left[X^2\right] - (EX)^2. \end{aligned}$$

Como se comentaba anteriormente, la interpretación de la varianza es la de un promedio que mide la distancia de los valores de la variable a la media de ésta. Si la varianza es pequeña, indica una alta concentración de los valores de la variable en torno a la media; y viceversa, si la varianza es grande, indica alta dispersión de los valores de la variable respecto de la media.

Ejemplo. Calculemos la varianza de una v.a. continua con función de densidad

$$f_X(x) = \begin{cases} \frac{1}{x_2 - x_1} \text{ si } x_1 \leq x \leq x_2 \\ 0 \text{ en otro caso} \end{cases}.$$

$$\begin{aligned} E\left[X^2\right] &= \int_{x_1}^{x_2} x^2 \cdot \frac{1}{x_2 - x_1} \cdot dx = \frac{1}{3} \frac{x_2^3 - x_1^3}{x_2 - x_1} \\ &= \frac{x_2^2 + x_1 x_2 + x_1^2}{3}. \end{aligned}$$

Anteriormente habíamos demostrado que

$$EX = \frac{x_1 + x_2}{2},$$

por tanto,

$$\begin{aligned} Var\left[X\right] &= E\left[X^2\right] - EX^2 \\ &= \frac{x_2^2 + x_1 x_2 + x_1^2}{3} - \frac{\left(x_1 + x_2\right)^2}{4} = \frac{\left(x_2 - x_1\right)^2}{12}. \end{aligned}$$

Nota. Estimaciones muestrales de media y varianza de una v.a.

Probablemente las mentes más despiertas ya se hayan planteado qué relación hay entre la media y la varianza de una v.a. (discreta o continua) y la media y la varianza de unos datos, definidas en el capítulo de Estadística Descriptiva.

La respuesta la veremos más adelante, pero podemos ir avanzando que la relación es parecida a la que se da entre los diagramas de barras y las funciones masa o entre los histogramas y las funciones de densidad. Es decir, si tenemos unos datos de una variable, en otras palabras, una muestra de una variable, la media y la varianza de la muestra serán aproximaciones de la media y la varianza de la variable aleatoria, aproximaciones que deben ser tanto mejores cuanto mayor sea el tamaño de la muestra.

Nota. Comportamiento de la media y la varianza frente a cambios de origen y escala.

Un cambio de origen de una variable consiste en sumar o restar una determinada cantidad a los valores de la variable, mientras que un cambio de escala supone multiplicar por un factor dichos valroes. En general, si X es una variable cualquiera, un cambio de origen y escala supone considerar $aX + b$.

Ya comentamos en el capítulo de Estadística Descriptiva el comportamiento de la media y la varianza muestral frente a estos cambios de origen y escala. Ahora nos referimos aquí al comportamiento de sus homólogos poblacionales. Este resultado es muy útil en la práctica y es válido tanto para variables continuas como para discretas. Concretamente, si X es una v.a. y $a,\ b \in \mathbb{R}$, entonces

$$\begin{aligned} E\left[aX + b\right] &= aE\left[X\right] + b \\ Var\left[aX + b\right] &= a^2 VarX \end{aligned}$$

Nota. Si tenemos una colección de variables aleatorias *independientes*, es decir, que son observadas sin que ninguna de ellas pueda influir sobre las otras, es muy útil plantearse en ocasiones por la media y la varianza de la suma de todas ellas.

Vamos a considerar las variables $X_1, ..., X_n$, que pueden ser discretas o continuas. Pues bien, se tiene que la media de la suma es la suma de las medias y que la varianza de la suma es la suma de las varianzas;

es decir,

$$E[X_1 + ... + X_n] = EX_1 + ... + EX_n$$
$$Var[X_1 + ... + X_n] = VarX_1 + ... + VarX_n$$

4.5. Modelos de distribuciones de probabilidad para variables continuas

Como en el caso de las variables discretas, vamos a describir a continuación los modelos de distribuciones de probabilidad más usuales para variables continuas.

De nuevo tenemos que insistir que la utilidad de estos modelos radica en que van a facilitarnos la manera en que se reparte la probabilidad de los valores de la variable.

4.5.1. Distribución uniforme (continua)

Se dice que una v.a. continua X que sólo puede tomar valores en el intervalo (x_1, x_2) sigue una **distribución uniforme entre** x_1 y x_2 (y se nota $X \to U(x_1, x_2)$) si su función de densidad es

$$f(x) = \begin{cases} \frac{1}{x_2 - x_1} \text{ si } x_1 < x < x_2 \\ 0 \text{ en otro caso} \end{cases}.$$

Sea $X \to U(x_1, x_2)$. Entonces

$$EX = \frac{x_1 + x_2}{2}$$
$$VarX = \frac{(x_2 - x_1)^2}{12}.$$

Caracterización de la distribución uniforme. Si X es una v.a. tal que dos intervalos cualesquiera entre x_1 y x_2 de la misma longitud, tienen la misma probabilidad, entonces $X \to U(x_1, x_2)$.

El ejemplo más habitual de esta variable es la variable uniforme en el intervalo $(0, 1)$; valores simulados de esta variable son los que se calculan con la orden **RND** de cualquier calculadora.

4.5.2. Distribución exponencial

Esta distribución suele ser modelo de aquellos fenómenos aleatorios que miden el tiempo que transcurre entre que ocurren dos sucesos. Por ejemplo, entre la puesta en marcha de una cierta componente y su fallo o el tiempo que transcurre entre dos llamadas consecutivas a una centralita.

Sea X una v.a. continua que puede tomar valores $x \geq 0$. Se dice que X sigue una **distribución exponencial de parámetro** λ (y se nota $X \to \exp(\mu)$) si su función de densidad

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \text{ si } x \geq 0 \\ 0 \text{ en otro caso} \end{cases}.$$

Obsérvese que su función de distribución es

$$F(x) = P[X \leq x] = \begin{cases} 1 - e^{-\lambda x} \text{ si } x \geq 0 \\ 0 \text{ en otro caso} \end{cases}.$$

Sea $X \to \exp(\lambda)$. Entonces,

$$EX = \frac{1}{\lambda}$$
$$VarX = \frac{1}{\lambda^2}.$$

Caracterización de la distribución exponencial. Sea $X \to P(\lambda)$ una v.a. discreta que cuenta el número de éxitos en un determinado periodo de tiempo. En ese caso, el tiempo que pasa entre dos éxitos consecutivos, T, es una v.a. que sigue una $\exp(\lambda)$.

Ejemplo. Un elemento radiactivo emite partículas según una variable de Poisson con un promedio de 15 partículas por minuto. En ese caso, el tiempo, T, que transcurre entre la emisión de una partícula y la siguiente sigue una distribución exponencial de parámetro $\lambda = 15$ partículas por minuto. Este modelo nos permite, por ejemplo, calcular la probabilidad de que entre partícula y partícula pasen más de 10 segundos, dado por

$$P[T > 10/60] = \int_{1/6}^{\infty} 15e^{-15t}dt = e^{-15/6}.$$

Ejemplo. Recordemos que habíamos comentado que la distribución de Poisson se solía utilizar en el contexto de las redes de comunicaciones como modelo para el número de solicitudes a un servidor por unidad de tiempo. Según esta caracterización que acabamos de ver, eso equivale a decir que el tiempo que pasa entre dos solicitudes a un servidor sigue una distribución exponencial.

Por ejemplo, supongamos que el número de conexiones a un servidor FTP sigue una distribución de Poisson de media 2.5 conexiones a la hora. En ese caso, podríamos preguntarnos cuál es la probabilidad de que pasen más de dos horas sin que se produzca ninguna conexión. Teniendo en cuenta que el tiempo entre conexiones seguiría una distribución exponencial de parámetro 2.5, esa probabilidad sería

$$P[T > 2] = \int_{2}^{\infty} 2.5e^{-2.5x}dx = e^{-5}$$

o bien

$$P[T > 2] = 1 - P[T \leq 2] = 1 - F_T(2) = 1 - \left(1 - e^{-2.5\times 2}\right) = e^{-5}.$$

Hay una interesante y curiosa propiedad de la distribución exponencial, conocida como *propiedad de no memoria*. Si X es una v.a. con distribución $exp(\lambda)$ y t y s son dos números positivos. Entonces:

$$P[X > t + s | X > s] = P[X > t]$$

La forma de demostrarlo es muy sencilla:

$$\begin{aligned} P[X > t + s|X > s] =& \frac{P[X > t + s \cap X > s]}{P[X > s]} = \frac{P[X > t + s]}{P[X > s]} \\ =& \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P[X > t] \end{aligned}$$

Vamos a tratar de entender la trascendencia de esta propiedad en el siguiente ejemplo.

Ejemplo. El tiempo de vida, T, de un circuito, sigue una distribución exponencial de media dos años. Calculemos la probabilidad de que un circuito dure más de tres años:

$$P[T > 3] = e^{-\frac{1}{2}3}$$

Supongamos que un circuito lleva 5 años funcionando, y que nos planteamos la probabilidad de que aún funcione 3 años más. Según la propiedad de no memoria, esa probabilidad es la misma que si el circuito acabara de comenzar a funcionar, es decir,

$$P[T > 3 + 5|T > 5] = P[T > 3] = e^{-\frac{1}{2}3}$$

Desde un punto de vista práctico, parece poco creible, porque entendemos que los 5 años previos de funcionamiento deben haber afectado a la fiabilidad del circuito, pero si creemos que la distribución del tiempo de vida de éste es exponencial, tenemos que asumir esta propiedad.

4.5.3. Distribución Gamma

Sea X una v.a. continua que puede tomar valores $x \geq 0$. Se dice que X sigue una **distribución Gamma** de parámetros a y λ (y se nota $X \to Gamma(a, \lambda)$) si su función de densidad es

$$f(x) = \frac{\lambda (\lambda x)^{a-1} e^{-\lambda x}}{\Gamma(a)} u(x),$$

donde $\Gamma(x) = \int_0^\infty s^{x-1} e^{-s} ds$ es la función gamma.

Obsérvese que en el caso en que $a = 1$ se tiene la distribución exponencial.

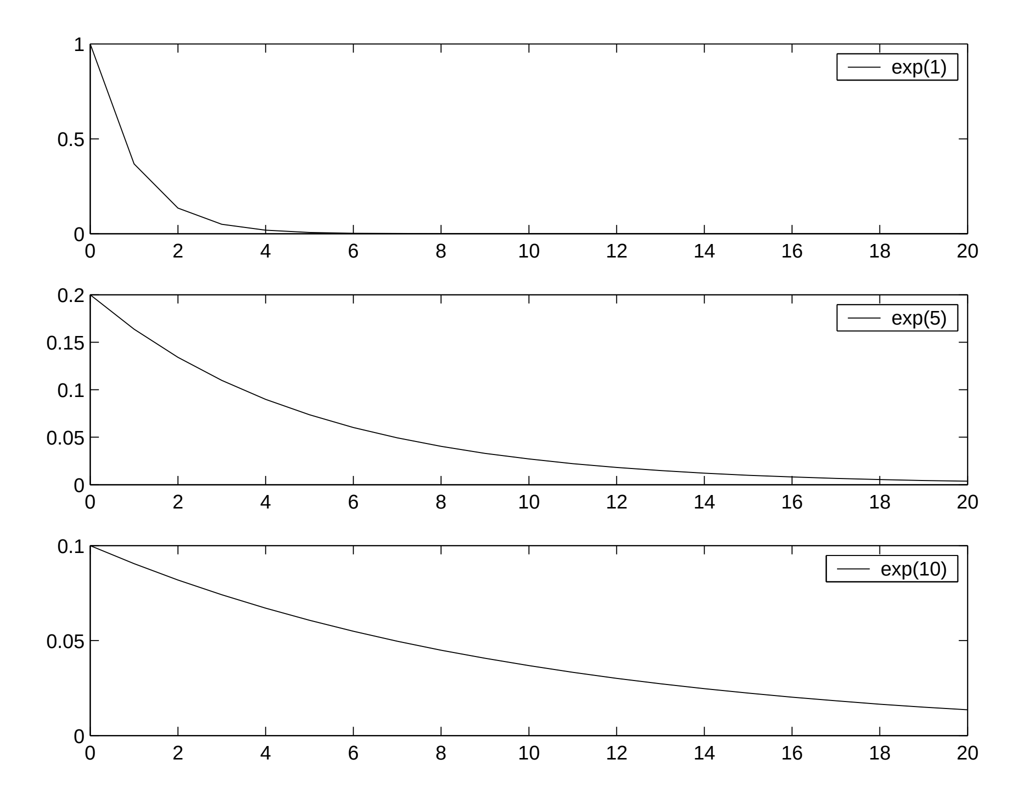


Figura 4.10: Funciones de densidad de distribuciones exponenciales.

En el contexto de las telecomunicaciones, hay un caso especialmente interesante. Si $a = n$, número natural, la distribución se denomina **Erlang.** Lo que la hace interesante es que esta distribución se utiliza como modelo del tiempo que pasa entre n llamadas telefónicas, por ejemplo.

Otro caso particular lo constituye la **distribución χ^2 con r grados de libertad**, que no es más que una $Gamma\left(\frac{r}{2}, \frac{1}{2}\right)$. Esta distribución se utiliza, por ejemplo, para evaluar la bondad del ajuste de una distribución teórica a unos datos, como veremos más adelante.

Sea $X \to Gamma\,(a, \lambda)$. Entonces

$$EX = \frac{a}{\lambda}$$
$$VarX = \frac{a}{\lambda^2}.$$

Caracterización de la distribución Gamma. Sea $X \to P\,(\lambda)$ una v.a. discreta que cuenta el número de éxitos en un determinado periodo de tiempo. En ese caso, el tiempo que pasa entre el $k-$ésimo éxito y el $k + r$, T, es una v.a. que sigue una $Gamma\,(r, \lambda)$. Dado que r es un entero, en realidad es una $Erlang\,(r, \lambda)$.

Caracterización de la distribución Gamma. Sean $X_1, ..., X_n$ v.a. independientes con distribución $\exp\,(\lambda)$. En ese caso, $X = \sum_{i=1}^{n} X_i$ sigue una $Gamma\,(n, \lambda)$. De nuevo obsérvese que el primer parámetro es un entero, luego se trata de una Erlang.

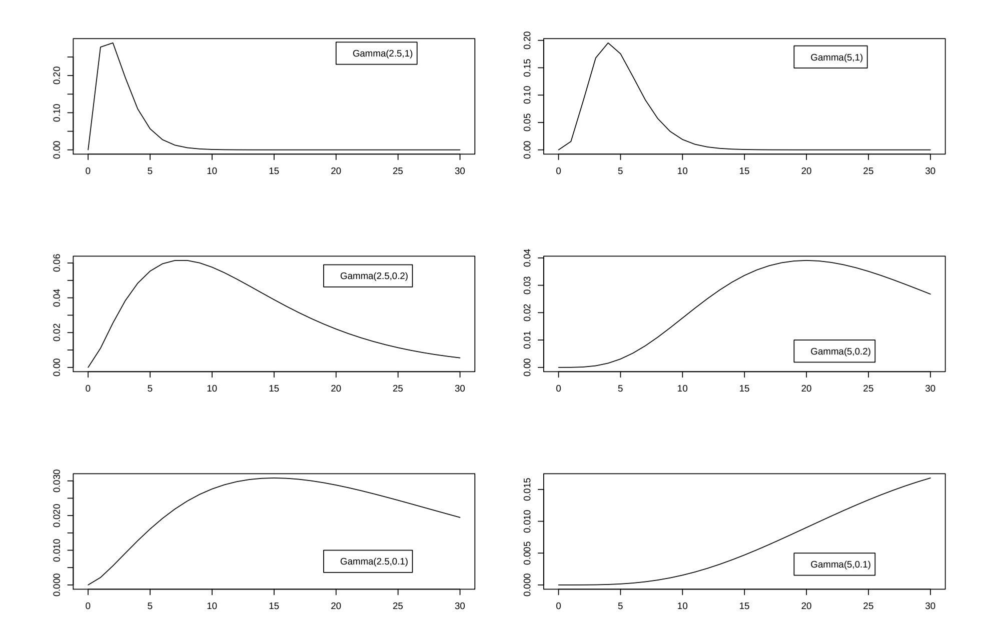


Figura 4.11: Funciones de densidad de distribuciones Gamma

4.5.4. Distribución normal

Sea X una v.a. continua que puede tomar cualquier valor real. Se dice que X sigue una **distribución normal o gaussiana, de parámetros** μ y σ (y se nota $X \to N(\mu, \sigma)$), si su función de densidad es

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \text{ para todo } x \in \mathbb{R}.$$

Obsérvese que es la única distribución que hemos visto hasta ahora que toma todos los valores entre $-\infty$ y $+\infty$.

Sea $X \to N(\mu, \sigma)$. Entonces

$$EX = \mu$$
$$VarX = \sigma^2.$$

El propio nombre de la distribución *normal* indica su frecuente uso en cualquier ámbito científico y tecnológico. Este uso tan extendido se justifica por la frecuencia o normalidad con la que ciertos fenómenos tienden a parecerse en su comportamiento a esta distribución, ya que muchas variables aleatorias continuas presentan una función de densidad cuya gráfica tiene forma de campana. Esto, a su vez, es debido a que hay muchas variables asociadas a fenómenos naturales cuyas características son compatibles con el modelo aleatorio que supone el modelo de la normal:

- Caracteres morfológicos de individuos (personas, animales, plantas, ...) de una especie (tallas, pesos, envergaduras, diámetros, perímetros, ...).

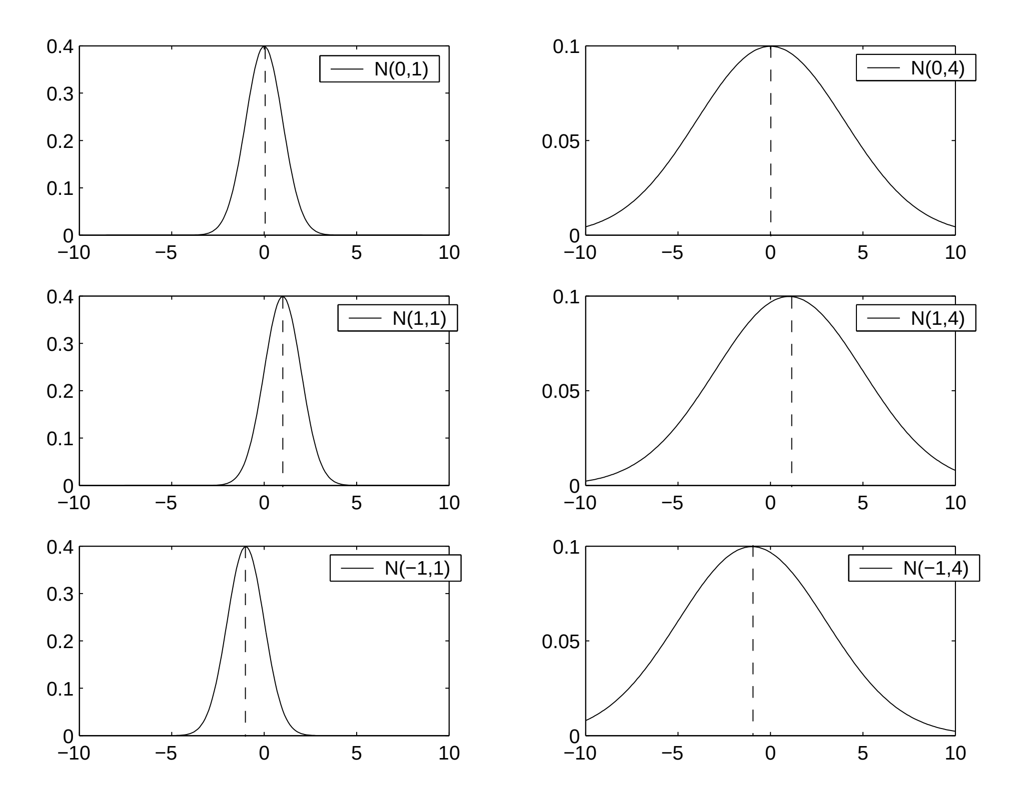


Figura 4.12: Funciones de densidad de la distribución normal

- Caracteres fisiológicos (efecto de una misma dosis de un fármaco, o de una misma cantidad de abono).
- Caracteres sociológicos (consumo de cierto producto por un mismo grupo de individuos, puntuaciones de examen...).
- Caracteres psicológicos (cociente intelectual, grado de adaptación a un medio, ...).
- Errores cometidos al medir ciertas magnitudes.
- Valores estadísticos muestrales, como por ejemplo la media.
- Otras distribuciones como la binomial o la de Poisson son aproximadas por la normal, ...

En general, como veremos enseguida, cualquier característica que se obtenga como suma de muchos factores independientes encuentra en la distribución normal un modelo adecuado.

Existe otra razón más pragmática para el uso tan extendido de la distribución normal: sus propiedades matemáticas son, como iremos viendo, casi inmejorables. Eso conduce a que casi siempre se trate de *forzar* al modelo normal como modelo para cualquier variable aleatoria, lo cual, en ocasiones puede conducir a errores importantes en las aplicaciones prácticas. Lo cierto es que también son frecuentes las aplicaciones en las que los datos no siguen una distribución normal. En ese caso puede ser relevante estudiar qué factores son los que provocan la pérdida de la normalidad y, en cualquier caso, pueden aplicarse técnicas estadísticas que no requieran de esa hipótesis.

Tipificación de la distribución normal. Sea $X \to N(\mu, \sigma)$. Entonces,

$$Z = \frac{X - \mu}{\sigma} \to N(0, 1),$$

propiedad que suele conocerse como *tipificación de la normal.*

Esta conocida propiedad tiene una aplicación práctica muy usual. Dadas las características de la densidad gaussiana, no es posible calcular probabilidades asociadas a la normal de forma exacta, ya que las integrales del tipo

$$\int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx$$

no pueden ser expresadas en términos de las funciones usuales, y sólo pueden calcularse por métodos numéricos. No obstante, existen tablas donde aparecen multitud de valores de la función de distribución de la distribución $N(0, 1)$ y a partir de ellos se pueden calcular otras tantas probabilidades, utilizando la propiedad de tipificación. Por ejemplo, si queremos calcular la probabilidad de que una variable $X \to N(\mu, \sigma)$ esté en el intervalo $[a, b]$, tenemos

$$P[a \leq X \leq b] = P\left[\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right] = F_Z\left(\frac{b-\mu}{\sigma}\right) - F_Z\left(\frac{a-\mu}{\sigma}\right),$$

donde $F_Z(\cdot)$ es la función de distribución de una variable $Z \to N(0, 1)$, que puede evaluarse mediante el uso de tablas. Vamos a verlo en un ejemplo.

Ejemplo. En el artículo "Índices de relación peso-talla como indicadores de masa muscular en el adulto del sexo masculino" de la revista **Revista Cubana Aliment. Nutr.** (1998;12(2):91-5) aparece un colectivo de varones con un peso cuya media y desviación estándar son, respectivamente, 65.6 y 11.7.

1. ¿Cómo podemos, mediante las tablas de la $N(0, 1)$, calcular, por ejemplo, la probabilidad de que uno de esos varones pese más de 76.25 kilos?

$$\begin{aligned} P[X > 76.25] &= P\left[\frac{X - 65.6}{11.7} > \frac{76.25 - 65.6}{11.7}\right] \\ &= P[Z > 0.91] = 1 - P[Z < 0.91] = 1 - 0.819 \end{aligned}$$

2. ¿Y la probabilidad de que pese menos de 60 kilos?

$$\begin{aligned} P[X < 60] &= P\left[\frac{X - 65.6}{11.7} < \frac{60 - 65.6}{11.7}\right] \\ &= P[Z < -0.48] = P[Z > 0.48] \\ &= 1 - P[Z < 0.48] = 1 - 0.684 \end{aligned}$$

3. ¿Y la probabilidad de que pese entre 60 y 76.25 kilos?

$$P[60 < X < 76.25] = P[X < 76.25] - P[X < 60] = 0.819 - (1 - 0.684)$$

Tabla 10: Función de distribución de la variable Normal(0,1)

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.0
0	0.500000	0.503989	0.507978	0.511966	0.515953	0.519939	0.523922	0.52790
0.1	0.539828	0.543795	0.547758	0.551717	0.555670	0.559618	0.563559	0.56749
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.60642
0.3	0.617911	0.621720	0.625516	0.629300	0.633072	0.636831	0.640576	0.64430
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.68082
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.71566
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.74857
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.77935
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802337	0.805105	0.80785
0.9	0.815940	0.818589	0.821214	0.823814	0.826391	0.828944	0.831472	0.83397
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.85769
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.87900

Figura 4.13: Búsqueda de probabilidades en la tabla de la $N(0,1)$. Valor de la probabilidad a la izquierda de 0.91

4. ¿Cuánto pesará aquel varón tal que un 5 % de varones de ese colectivo pesan más que él? Es decir, ¿cuál será el valor de x tal que $P[X > x] = 0.05$ o, equivalentemente, $P[X < x] = 0.95$. Dado que

$$P[X < x] = P\left[\frac{X - 65.6}{11.7} < \frac{x - 65.6}{11.7}\right] = P\left[Z < \frac{x - 65.6}{11.7}\right]$$

tan sólo tenemos que buscar el valor $z = \frac{x-65.6}{11.7}$ tal que $P[Z < z] = 0.95$, 1.645 (aproximadamente), en cuyo caso, $x = 65.6 + 11.7 \times 1.645$.

Tabla 10: Función de distribución de la variable Normal(0,1)

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0	0.500000	0.503989	0.507978	0.511966	0.515953	0.519939	0.523922	0.52790
0.1	0.539828	0.543795	0.547758	0.551717	0.555670	0.559618	0.563559	0.56749
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.60642
0.3	0.617911	0.621720	0.625516	0.629300	0.633072	0.636831	0.640576	0.64430
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.68082
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.71566
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.74857
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.77935
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802337	0.805105	0.80785
0.9	0.815940	0.818589	0.821214	0.823814	0.826391	0.828944	0.831472	0.83397
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.85769
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.87900
1.2	0.884930	0.886861	0.888768	0.890651	0.892512	0.894350	0.896165	0.89795
1.3	0.903200	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.91465
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.92921
1.5	0.933193	0.934478	0.935745	0.936992	0.938220	0.939429	0.940620	0.94179
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.95254
1.7	0.955435	0.956367	0.957284	0.958185	0.959070	0.959941	0.960796	0.96163

Figura 4.14: Búsqueda de valores z en la tabla de la $N(0,1)$. Valor de Z que deja a la derecha una probabilidad de 0.95

Teorema Central del Límite. Sean $X_1, ..., X_N$ v.a. independientes, todas ellas con la misma distribución de probabilidad, distribución de media μ_X y desviación típica σ_X. En ese caso, la suma de estas variables sigue aproximadamente una distribución normal cuando N es elevado, es decir,

$$\sum_{i=1}^{N} X_i \approx N\left(N\mu_X, \sqrt{N}\sigma_X\right).$$

Tipificando, podemos reenunciar el Teorema Central del Límite diciendo que

$$\frac{\sum_{i=1}^{N} X_i - N\mu_X}{\sqrt{N}\sigma_X} \approx N\left(0,1\right).$$

Este teorema es el que proporciona una justificación matemática del porqué la distribución gaussiana es un modelo adecuado para un gran número de fenómenos reales en donde la v.a. observada en un momento dado es el resultado de sumar un gran número de sucesos aleatorios elementales.

Ejemplo. Consideremos $X_1, ..., X_N$ variables independientes con distribución $U[0,1]$. Según el teorema central del límite, $\sum_{i=1}^{N} X_i \approx N\left(0.5N, \sqrt{\frac{N}{12}}\right)$. Para poner este resultado de manifiesto se ha realizado el siguiente experimento:

Para $N = 1, 2, 5$ y 10, se ha simulado una muestra de 10000 datos de $\sum_{i=1}^{N} X_i$, dibujando su histograma

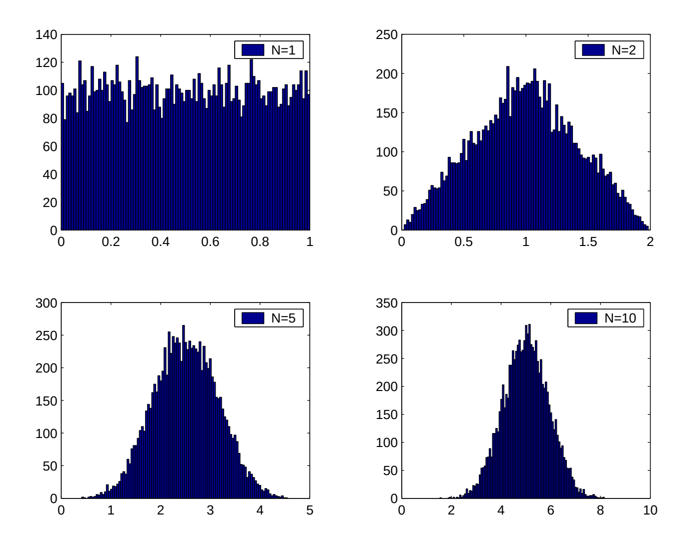


Figura 4.15: Ilustración del Teorema Central del Límite.

en cada caso. Estos histogramas aparecen en la Figura 4.15. En ella se pone de manifiesto como según N crece, el histograma se va pareciendo cada vez más a una densidad gaussiana.

Ejemplo. Supongamos que estamos realizando un examen de 150 preguntas, cada una de ellas con una puntuación de 1 punto y que en función de cómo hemos estudiado, consideramos que la probabilidad de contestar acertadamente una pregunta cualquiera es de 0.7. Démonos cuenta que el resultado de una pregunta cualquiera sigue una distribución $B(1, 0.7)$, cuya media es $1 \times 0.7 = 0.7$ y cuya varianza es $1 \times 0.7 \times (1 - 0.7) = 0.21$.

Por su parte, el resultado final de la prueba será la suma de las 150 puntuaciones. Podríamos ver este resultado según una $B(150, 0.7)$, pero los cálculos serían muy tediosos debido a los factoriales de la función masa de la distribución binomial. En este caso, merece la pena que utilicemos el Teorema Central del Límite, según el cuál el resultado final, X, seguiría aproximadamente una distribución

$$N\left(150 \times 0.7, \sqrt{150 \times 0.21}\right),$$

es decir, $X \to N(105, 5.612)$. Así, si por ejemplo, nos planteamos cuál es la probabilidad de aprobar, ésta será

$$P[X > 75] = P[Z > -0.952] = 0.830.$$

Esta aplicación se conoce, en general, como *aproximación normal de la binomial.*

Enunciando el Teorema Central del Límite en términos de la media, $\bar{X}$, de las variables $X_1, ..., X_N$, podemos decir que si N es grande,

$$\bar{X} \approx N(\mu, \sigma/\sqrt{N})$$

Ejemplo. Un ingeniero diseña un aparato de medida que realiza una aproximación más imprecisa que el aparato tradicional pero mucho más barata. Para reducir el margen de error de la medida realizada, el ingeniero propondrá que se realicen un número determinado de medidas sobre el mismo objeto y que se considere la media de estas medidas como valor final de la medida del objeto.

Inicialmente, el ingeniero hace una valoración que le lleva a concluir que el aparato está bien calibrado, es decir, que la media de la medida del aparato coincide con la medida real, y que la desviación típica de las medidas del aparato es igual a 0.75.

¿Cuántas medidas debe proponer el ingeniero para que el error de medida sea inferior a 0.1 con un 95 % de probabilidad?

Empecemos considerando que cada medida, X_i, tiene como media el verdadero valor de la medida del objeto, x_0, y desviación típica 0.75. Por su parte, la medida final será $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, donde realmente nos interesa conocer el valor de n. Para ello, tengamos en cuenta que se nos pide que

$$P\left[\left|\bar{X} - x_0\right| < 0.1\right] \geq 0.95.$$

y que, considerando el Teorema Central del Límite, $\bar{X} \to N\left(x_0, \frac{0.75}{\sqrt{n}}\right)$. Por su parte,

$$\begin{aligned} P\left[\left|\bar{X} - x_0\right| < 0.1\right] &= P\left[x_0 - 0.1 < \bar{X} < x_0 + 0.1\right] = P\left[-\frac{0.1\sqrt{n}}{0.75} < Z < \frac{0.1\sqrt{n}}{0.75}\right] \\ &= 1 - 2 \times \left(1 - P\left[Z < \frac{0.1\sqrt{n}}{0.75}\right]\right). \end{aligned}$$

Si queremos que $P\left[\left|\bar{X} - x_0\right| < 0.1\right] \geq 0.95$, entonces $P\left[Z < \frac{0.1\sqrt{n}}{0.75}\right] \geq 0.975$, de donde $\frac{0.1\sqrt{n}}{0.75} \geq 1.96$ y entonces, $n \geq 216.09$.

Como conclusión, más le vale al ingeniero disminuir la desviación típica del aparato de medida.

4.6. Cuantiles de una distribución. Aplicaciones

Para acabar el tema vamos a ver una de las aplicaciones más sencillas pero a la vez más útiles de los modelos de probabilidad. Debo decir que son numerosas las ocasiones que desde distintos ambientes científicos y de la Ingeniería he asesorado a profesionales con respecto a cuestiones que tienen que ver con lo que esta sección analiza. Los ejemplos que vamos a considerar son, *grosso modo*, síntesis de ellas.

Concretamente, vamos a comenzar definiendo el **cuantil** p ($p \in [0, 1]$) de una distribución de probabilidad de una v.a. X. Sea ésta discreta o continua, denominemos $f(x)$ a su función masa o de densidad.

Se define el cuantil p, Q_p de su distribución como el primer valor, x, de la variable tal que $P[X \leq x] \geq p$:

- Si la variable es discreta, Q_p será, por tanto, el primer valor tal que

$$\sum_{x_i \leq x} f(x) \geq p.$$

Nótese que, al ser la variable discreta, puede que no logremos obtener una igualdad del tipo $\sum_{x_i \leq x} f(x) = p$.

- Si la variable es continua, Q_p sí puede obtenerse como el valor x tal que

$$\int_{-\infty}^{x} f(t)\, dt = p,$$

o lo que es lo mismo, como el valor x tal que $F(x) = p$, siendo F la función de distribución de la variable.

Es muy frecuente que la probabilidad p a la que se asocia un cuantil se exprese en porcentaje. En ese caso, los cuantiles también se pueden llamar **percentiles**. Por ejemplo, el cuantil 0.5 es el percentil 50, la mediana.

Desde luego, lo más importante es que interpretemos qué significa el cuantil p de una v.a. Como en Estadística Descriptiva, se refiere al valor de la variable que deja por debajo de sí una proporción p de valores de la variable. Entonces, si un valor concreto corresponde con un cuantil *alto*, podemos decir que realmente es un valor *alto* dentro de la distribución de probabilidad de la variable, y viceversa. Vamos a tratar de aclararlo con algunos ejemplos.

4.6.1. La bombilla de bajo consumo marca ANTE

En el capítulo de introducción comentábamos las especificaciones técnicas que aparecían en el envoltorio de una bombilla de 14W de la marca ANTE, entre las que se decía que tenía una duración de 8 años. Eso contradice nuestra sensación de que este tipo de lámparas duran mucho menos y, en cualquier caso, es una simplificación inadmisible, porque es evidente que la duración de la bombilla es una variable sujeta a incertidumbre, es decir, una variable aleatoria.

Vamos a hacer un par de asunciones. En primer lugar, es probable que lo que quisieran decir en el envoltorio es que la **duración media** es de 8 años (lo cuál, por cierto, también podría ser objeto de controversia). En segundo lugar, dado que tenemos que proponer un modelo de distribución de probabilidad para la duración de la lámpara, vamos a considerar el más sencillo que suele emplearse en este tipo de aplicaciones: la distribución exponencial. Esta hipótesis también podría ser discutida, pero otros modelos más complejos, como la distribución Weibull, complicarían bastante nuestros cálculos que, por otra parte, tienen sólo fines ilustrativos.

Por tanto, vamos a suponer que la duración de la bombilla es una variable aleatoria, D, con distribución exponencial de media 8 años y, por tanto, con parámetro $\lambda = 1/8$. Ahora que ya tenemos un modelo probabilístico podemos plantearnos muchas cosas:

- ¿Es muy probable que la lámpara alcance su vida media?

$$P[D > 8] = \int_{8}^{\infty} \frac{1}{8} e^{-\frac{x}{8}} dx = e^{-8/8} = 0.3678794.$$

Obsérvese que eso es algo que ocurrirá con cualquier exponencial: la probabilidad de que se supere la media es sólo del 36.79 %. Dicho de otra forma, la media es el percentil 63 aproximadamente, lo que implica que sólo el 37 % aproximadamente de las lámparas superan su vida media... ¿sorprendente?

- ¿Y cuál es el valor que superan el 50 % de las lámparas? Se trata de la mediana, $Me = F^{-1}(0.5)$, donde $F()$ es la función de distribución. Por tanto, la mediana es la solución de la ecuación

$$1 - e^{-\lambda Me} = 0.5,$$

  que resulta ser $Me = \frac{log0.5}{-\lambda} = 8 \times log2 = 5.545177$. Luego, visto de otra forma, el 50 % de las lámparas se rompen antes de 5.545 años.

Para terminar, animo a los lectores interesados a que busquen información sobre el cómputo de la vida media de este tipo de lámparas, basado en la realización de pruebas aceleradas sobre una muestra (bastante reducida, por cierto) de lámparas.

4.6.2. Las visitas al pediatra de los padres preocupados

Los que tenemos hijos pequeños observamos con cierta ansiedad la evolución de su peso y su altura. Cuando vamos al pediatra, éste pesa y mide al bebé y, obviamente, te dice *cómo está*. Pero el problema es que no basta con que me diga cuánto pesa y mide mi hijo o mi hija, sino que me diga cuánto pesa y cuánto mide en relación con los niños o niñas de su misma edad. En esa cuestión es dónde entran los percentiles.

En este caso jugamos con la ventaja de que se han hecho multitud de estudios previos que determinan que tanto el peso como la altura son variables que siguen una distribución normal. Más aún, se han determinado las medias y las desviaciones típicas de niños y niñas desde los 0 meses hasta la edad adulta.

Vamos a ponernos en una situación concreta, centrándonos en el peso. Tengo un hijo de tres meses que pesa 5.6 kilos. La pregunta es *¿está gordo? ¿es bajito?* En cualquier caso, *cómo de gordo o de bajito*. El pediatra sabe por estudios previos[2] que el peso de niños de tres meses es una $N(6, 1.2)$. Lo que se plantea es en qué posición se sitúa el peso de mi hijo, 5.6 kilos, dentro de esa distribución. Si X es el peso, dado que

$$P[X \leq 5.6] = 0.369,$$

el pediatra me dirá que mi hijo está en el percentil 37, lo que quiere decir que es un pelín bajo de peso, pero dentro de niveles razonables.

[2]Fuente: http://www.familia.cl/salud/curvas_de_crecimiento/curvas_de_crecimiento.htm

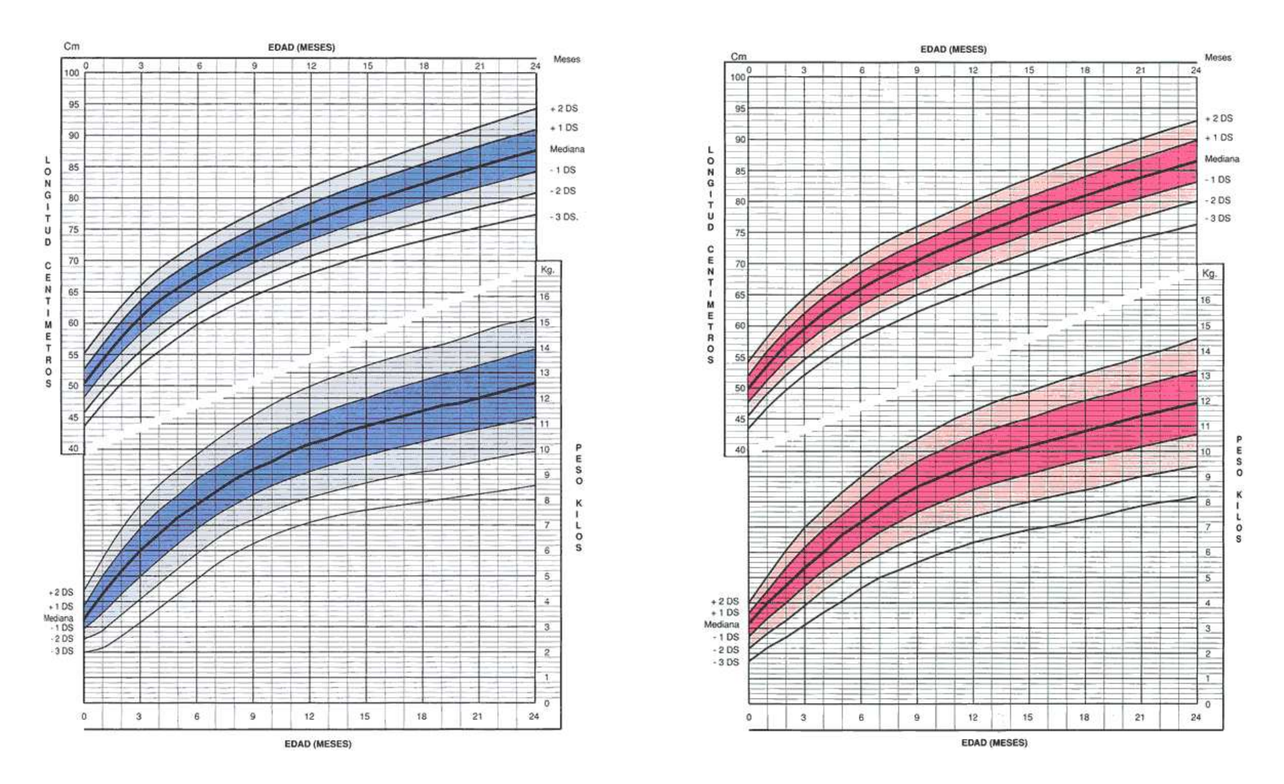


Figura 4.16: Curvas de crecimiento de 0 a 24 meses.

Capítulo 5

Variables aleatorias con distribución conjunta

> El matrimonio es la principal causa de divorcio.
>
> Groucho Marx

Resumen. En el estudio de las variables aleatorias hemos pasado por alto el hecho de que un conjunto de dos o más variables puede verse afectado por una serie de relaciones entre ellas. El análisis desde el punto de vista estadístico de estas relaciones es el objetivo de este capítulo. Como caso especial, describiremos de forma detallada el modelo que para estas relaciones proporciona la distribución normal multivariante

Palabras clave: distribución conjunta, distribución marginal, distribución condicionada, covarianza, coeficiente de correlación, normal multivariante.

5.1. Introducción

El mundo real está repleto de relaciones a todos los niveles. Nosotros, por razones obvias, estaremos interesados principalmente en las relaciones que afectan a variables que describen fenómenos propios del ambiente científico-tecnológico. Estas relaciones pueden tener muy diversas tipologias. Por ejemplo, podríamos pensar en relaciones causa-efecto, como la que, por ejemplo, explicaría que una página Web tenga un tamaño considerable *debido* a que lleva incrustado varios archivos de vídeo y audio, o la que se establece entre la edad en años de un vestigio y su contenido en carbono 14[1]. Pero no sólo tendremos relaciones causa-efecto: por ejemplo, sabemos que el peso y la estatura de un ser humano son variables muy relacionadas, hasta el punto que no podemos decir que una persona este obesa sólo con saber su peso, sino que debemos valorarlo *en relación a* su estatura.

Por otra parte, cuando un fenómeno es determinístico y está bien estudiado, las relaciones entre variables son leyes más o menos sencillas, pero, en cualquier caso, son inmutables. Por ejemplo,

$$densidad = \frac{masa}{vol.}.$$

[1] Relación que, por cierto, sabemos que permite la datación del vestigio.

Pero, ¿qué ocurre cuando el fenómeno es aleatorio? Las variables en ese caso son aleatorias y las relaciones que se puedan dar entre ellas no siempre tienen por qué obedecer a una ley objetiva e inamovible. Por ejemplo, todos somos conscientes de que, como decíamos, existe una relación entre el peso y la altura de una persona, pero no existe una *razón de conversión* capaz de calcular el peso exacto de alguien a partir de su altura. Es evidente que el tiempo de descarga de una página web estará relacionado con el tamaño de los archivos que la configuran, pero ¿cómo de *evidente*? y ¿de qué forma es esa relación? Ambas preguntas tratarán de ser contestadas a lo largo de este capítulo.

Sean $X_1, ..., X_N$ variables aleatorias. El vector ordenado

$$\begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}$$

es un **vector aleatorio de dimensión** N.

Hablaremos de **vectores aleatorios continuos** o **vectores aleatorios discretos** cuando cada una de sus variables sean continuas o discretas, respectivamente. Podrían darse *vectores mixtos*, pero su tratamiento estadístico no nos interesa por ahora.

Ejemplo. Consideremos el valor de una señal analógica que depende del tiempo, $x(t)$. En esta notación, entendemos que el valor de la señal podría ser distinto en cada instante de tiempo t. Es muy frecuente que la señal se observe realmente contaminada por un ruido aleatorio que también dependerá del tiempo, $N(t)$. En ese caso, si observamos la señal en los instantes $t_1, ..., t_N$, el vector

$$\begin{pmatrix} x(t_1) + N(t_1) \\ \vdots \\ x(t_n) + N(t_n) \end{pmatrix}$$

es un vector aleatorio.

Ejemplo. Se estudia el tiempo que un usuario de Internet dedica a ver una página WEB (T) en relación con variables como la cantidad de texto que contiene (Tx), el número de imágenes (I) y animaciones Flash (F) de la página. Entonces, el vector

$$\begin{pmatrix} T \\ Tx \\ I \\ F \end{pmatrix}$$

es un vector aleatorio.

Ejemplo. Se contabiliza la duración de las llamadas telefónicas a una centralita. Para cada conjunto de n-usuarios de la centralita, cada uno de ellos ocupa un tiempo T_i en su llamada. En ese caso, el vector

$$\left(\begin{array}{c} T_1 \\ \vdots \\ T_n \end{array} \right)$$

es un vector aleatorio.

5.2. Distribuciones conjunta, marginal y condicionada

El principal objetivo a abordar en el tema es cómo medir la incertidumbre asociada a los sucesos que describe un vector aleatorio. Ya vimos que en el caso de una variable aleatoria se trataba de hacerlo a partir de la función masa o la función de densidad. Ahora, como vamos a ver, es algo más complejo.

5.2.1. Distribución conjunta

La **distribución conjunta de probabilidad** de un vector aleatorio es, esencialmente, la manera en que se reparte la probabilidad entre todos los posibles resultados del vector. Para describirla vamos a definir los conceptos de función de densidad o función masa análogos a los asociados a una variable aleatoria.

Sea $(X_1, ..., X_N)$ un vector aleatorio discreto. Entonces, se define su **función masa conjunta** como

$$f_{X_1,...,X_N}(x_1, ..., x_N) = P[X = x_1, ..., X_N = x_N].$$

Por su parte, si $(X_1, ..., X_N)$ es un vector aleatorio continuo, entonces, su **función de densidad conjunta** es una función tal que

$$P\left[(X_1, ..., X_N) \in A \subset \mathbf{R}^N\right] = \int \underset{A\subset\mathbf{R}^N}{...} \int f_{X_1,...,X_N}(x_1, ..., x_N)\, dx_1...dx_N$$

Ejemplo. Consideremos un vector aleatorio bidimensional,$(X, Y)'$, que tiene densidad conjunta

$$f_{X,Y}(x, y) = \begin{cases} ce^{-x-y} \text{ si } 0 < y < x \\ 0 \text{ en otro caso} \end{cases}.$$

En primer lugar, podemos calcular la constante c teniendo en cuenta que

$$\int_{\mathbf{R}^2} f_{X,Y}(x, y)\, dxdy = 1.$$

Por ello,

$$1 = \int_0^\infty \left(\int_0^x ce^{-x} e^{-y} dy \right) dx = \int_0^\infty ce^{-x} \left(1 - e^{-x}\right) dx = \frac{c}{2},$$

de donde $c = 2$.

En segundo lugar, por ejemplo, calculemos

$$\begin{aligned} P[X + Y \leq 1] &= \int_0^1 \int_y^{1-y} 2e^{-x} e^{-y} dx dy \\ &= \int_0^1 2e^{-y} \left[e^{-y} - e^{-(1-y)} \right] dy \\ &= \frac{-1 - 2e + e^2}{e^2}. \end{aligned}$$

(ver Figura 5.1)

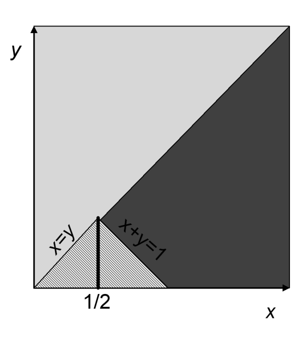


Figura 5.1: Región del plano donde se calcula la probabilidad.

Ejemplo. Consideremos dos variables, X e Y, que tienen densidad conjunta

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{15} \text{ si } 0 \leq x \leq 3, \ 0 \leq y \leq 5 \\ 0 \text{ en otro caso} \end{cases}.$$

Esta densidad constante en el rectángulo definido indica que la distribución de probabilidad es uniforme en dicho rectángulo. Vamos a calcular la probabilidad de que Y sea mayor que X (ver Figura 5.2)

$$\begin{aligned} P[Y > X] &= \int_0^3 \left(\int_x^5 \frac{1}{15} dy \right) dx \\ &= \int_0^3 \frac{5 - x}{15} dx \\ &= \frac{x}{3} - \frac{x^2}{30} |_0^3 = \frac{7}{10}. \end{aligned}$$

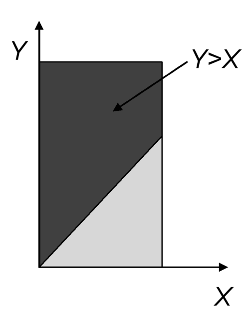


Figura 5.2: Región del plano donde se calcula la probabilidad.

5.2.2. Distribuciones marginales

Una vez que somos capaces de describir la distribución de probabilidad de un vector aleatorio mediante su función masa o su función de densidad conjunta, surge un nuevo problema: qué ocurre si deseamos conocer la distribución de probabilidad de una o más variables del vector, no del vector en su conjunto. Esa distribución de una o más variables de un vector se conoce como **distribución marginal**.

Sea $(X_1, ..., X_N)'$ un vector aleatorio y $(X_{i_1}, ..., X_{i_k})$ un subvector de variables suyo. En ese caso:
Si el vector es continuo,

$$f_{X_{i_1},...,X_{i_k}}(x_{i_1}, ..., x_{i_k}) = \int_{x_j \notin (x_{i_1},...,x_{i_k})} ... \int f_{X_1,...X_N}(x_1, ..., x_n) \prod_{x_j \notin (x_{i_1},...,x_{i_k})} dx_j.$$

Si el vector es discreto,

$$f_{X_{i_1},...,X_{i_k}}(x_{i_1}, ..., x_{i_k}) = \sum_{x_j \notin (x_{i_1},...,x_{i_k})} f_{X_1,...X_N}(x_1, ..., x_n).$$

Ejemplo. Sea el vector bidimensional (X, Y) con función de densidad conjunta $f_{X,Y}(x, y) = x \cdot e^{-x(y+1)}$ para $x, y > 0$.

La función de densidad marginal de X,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy = \int_0^{\infty} xe^{-x(y+1)} dy = e^{-x}$$

para $x > 0$.

Análogamente, la función de densidad marginal de Y,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \cdot dx = \int_0^{\infty} xe^{-x(y+1)} dx = \frac{1}{(1+y)^2}$$

para $y > 0$.

Ejemplo. Consideremos dos variables discretas, Q y G, cuya función masa, $f_{Q,G}(q,g)$, viene dada por

$f_{Q,G}(q,g)$	$g=0$	$g=1$	$g=2$	$g=3$
$q=0$	0.06	0.18	0.24	0.12
$q=1$	0.04	0.12	0.16	0.08

Sus marginales respectivas son:

$$\begin{aligned} f_Q(q) &= \sum_g f_{Q,G}(q,g) \\ &= \begin{cases} 0.06+0.18+0.24+0.12 \text{ si } q=0 \\ 0.04+0.12+0.16+0.08 \text{ si } q=1 \end{cases} \\ &= \begin{cases} 0.6 \text{ si } q=0 \\ 0.4 \text{ si } q=1 \end{cases} \end{aligned}$$

y

$$f_G(g) = \begin{cases} 0.06+0.04 & \text{si } g=0 \\ 0.18+0.12 & \text{si } g=1 \\ 0.24+0.16 & \text{si } g=2 \\ 0.12+0.08 & \text{si } g=3 \end{cases}$$

Ejemplo. En un ejemplo anterior considerábamos dos variables X e Y que tienen densidad conjunta

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{15} \text{ si } 0 \le x \le 3,\ 0 \le y \le 5 \\ 0 \text{ en otro caso} \end{cases}.$$

Vamos a calcular sus densidades marginales:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \\ &= \begin{cases} \int_0^5 \frac{1}{15}dy \text{ si } 0 \le x \le 3 \\ 0 \text{ en otro caso} \end{cases} \\ &= \begin{cases} \frac{1}{3} \text{ si } 0 \le x \le 3 \\ 0 \text{ en otro caso} \end{cases} \end{aligned}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$$
$$= \begin{cases} \int_0^3 \frac{1}{15}dx \text{ si } 0 \leq y \leq 5 \\ 0 \text{ en otro caso} \end{cases}$$
$$= \begin{cases} \frac{1}{5} \text{ si } 0 \leq y \leq 5 \\ 0 \text{ en otro caso} \end{cases}.$$

Por tanto, ambas marginales corresponden a sendas densidades uniformes.

Ejemplo. La densidad conjunta de X e Y es

$$f_{X,Y}(x,y) = \begin{cases} 2x \text{ si } 0 \leq x \leq 1,\ \ |y| < x^2 \\ 0 \text{ en otro caso} \end{cases}.$$

Calculemos ambas marginales:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$$
$$= \begin{cases} \int_{-x^2}^{x^2} 2xdy \text{ si } 0 \leq x \leq 1 \\ 0 \text{ en otro caso} \end{cases}$$
$$= \begin{cases} 4x^3 \text{ si } 0 \leq x \leq 1 \\ 0 \text{ en otro caso} \end{cases}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$$
$$= \begin{cases} \int_{\sqrt{|y|}}^{1} 2xdx \text{ si } -1 \leq y \leq 1 \\ 0 \text{ en otro caso} \end{cases}$$
$$= \begin{cases} 1-|y| \text{ si } -1 \leq y \leq 1 \\ 0 \text{ en otro caso} \end{cases}.$$

5.2.3. Distribuciones condicionadas

Si tenemos un vector $X = (X_1, ..., X_N)'$, podemos considerar la distribución de probabilidad de un vector formado por un subconjunto de variables de X, $(X_{i_1}, ..., X_{i_k})'$, condicionada al hecho de que se han dado determinados valores en otro subconjunto de variables de X, $X_{j_1} = x_{j_1}, ..., X_{j_l} = x_{j_l}$.

Esta distribución vendrá caracterizada por su función masa o su función de densidad **condicionadas**, según sea el vector discreto o continuo, y tendrá la expresión

$$f_{X_{i_1},...,X_{i_k}} |_{X_{j_1}=x_{j_1},...,X_{j_l}=x_{j_l}} (x_{i_1}, ..., x_{i_k}) = \frac{f_{X_{i_1},...,X_{i_k},X_{j_1},...,X_{j_l}} (x_{i_1}, ..., x_{i_k}, x_{j_1}, ..., x_{j_l})}{f_{X_{j_1},...,X_{j_l}} (x_{j_1}, ..., x_{j_l})},$$

donde $f_{X_{i_1},...,X_{i_k},X_{j_1},...,X_{j_l}} (x_{i_1}, ..., x_{i_k}, x_{j_1}, ..., x_{j_l})$ es la función masa o la función de densidad conjunta de las variables $X_{i_1}, ..., X_{i_k}, X_{j_1}, ..., X_{j_l}$ y $f_{X_{j_1},...,X_{j_l}} (x_{j_1}, ..., x_{j_l})$ es la función masa o la función de densidad conjunta de las variables $X_{j_1}, ..., X_{j_l}$.

En el caso más habitual en el que el vector tenga dimensión dos, tenemos la densidad o la función masa de X condicionada a $Y = y$,

$$f_{X|Y=y} (x) = \frac{f_{X,Y} (x, y)}{f_Y (y)}$$

o la densidad o la función masa de Y condicionada a $X = x$,

$$f_{Y|X=x} (y) = \frac{f_{X,Y} (x, y)}{f_X (x)}.$$

Ejemplo. Sean las variables X e Y con la función masa conjunta siguiente:

y\x	0	1	2
0	3/28	9/28	3/28
1	3/14	3/14	0
2	1/28	0	0

Las marginales son

$$f_X (x) = \begin{cases} \frac{3}{28} + \frac{3}{14} + \frac{1}{28} \text{ si } x = 0 \\ \frac{9}{28} + \frac{3}{14} + 0 \text{ si } x = 1 \\ \frac{3}{28} + 0 + 0 \text{ si } x = 2 \end{cases}$$

y

$$f_Y (y) = \begin{cases} \frac{3}{28} + \frac{9}{28} + \frac{3}{28} \text{ si } y = 0 \\ \frac{3}{14} + \frac{3}{14} + 0 \text{ si } y = 1 \\ \frac{1}{28} + 0 + 0 \text{ si } y = 2 \end{cases}$$

Como ejemplos de las condicionadas (hay 6 en total) calculemos la función masa de X condicionada a $Y = 1$ y la de Y condicionada a $X = 1$.

$$f_{X|Y=1} (x) = \begin{cases} \frac{\frac{3}{14}}{\frac{6}{14}} \text{ si } x = 0 \\ \frac{\frac{3}{14}}{\frac{6}{14}} \text{ si } x = 1 \\ \frac{0}{\frac{6}{14}} \text{ si } x = 2 \end{cases}.$$

$$f_{Y|X=1} (y) = \begin{cases} \frac{\frac{9}{28}}{\frac{15}{28}} \text{ si } y = 0 \\ \frac{\frac{3}{14}}{\frac{15}{28}} \text{ si } x = 1 \\ \frac{0}{\frac{15}{28}} \text{ si } x = 2 \end{cases}.$$

Como es evidente, una vez que tenemos caracterizada la distribución condicionada de una variable aleatoria al valor de otra, cualquier característica de dicha distribución, como la media o la varianza, puede calcularse a partir de su función masa o su función de densidad.

Ejemplo. Tal y como planteábamos al comienzo del capítulo, supongamos que la posición (X, Y) de un teléfono móvil que recibe cobertura de una antena de telefonía se encuentra dentro de un círculo de radio r alrededor de esa antena, que supondremos sin pérdida de generalidad que se encuentra en el origen del plano. Vamos a suponer que esa posición es *completamente al azar* dentro del círculo. Eso equivale a considerar que la densidad conjunta debe ser constante en el círculo; para que su integral sea la unidad, es evidente que

$$f_{X,Y}(x, y) = \frac{1}{\pi r^2}$$

si $x^2 + y^2 \leq r^2$ y cero en cualquier punto fuera del círculo. Vamos a ver qué podemos averiguar sobre las coordenadas X e Y por separado (marginales) y sobre cómo afectan la una a la otra (condicionadas).

En primer lugar,

$$f_X(x) = \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} \frac{1}{\pi r^2} dy = \frac{2\sqrt{r^2 - x^2}}{\pi r^2}$$

si $-r < x < r$. La marginal de Y es análoga,

$$f_Y(y) = \frac{2\sqrt{r^2 - y^2}}{\pi r^2}$$

si $-r < y < r$. Está claro que para cada coordenada por separado, los puntos *más densos*, *más probables*, son los cercanos al origen, que es donde se da el máximo de ambas funciones.

Ahora supongamos que conocemos una de las coordenadas y veamos qué podemos decir sobre la otra:

$$f_{X|Y=y_0}(x) = \frac{f_{X,Y}(x, y_0)}{f_Y(y_0)} = \frac{1}{2\sqrt{r^2 - y_0^2}}$$

si $-\sqrt{r^2 - y_0^2} < x < \sqrt{r^2 - y_0^2}$. Análogamente,

$$f_{Y|X=x_0}(y) = \frac{f_{X,Y}(x_0, y)}{f_X(x_0)} = \frac{1}{2\sqrt{r^2 - x_0^2}}$$

si $-\sqrt{r^2 - x_0^2} < y < \sqrt{r^2 - x_0^2}$. Si nos damos cuenta, ambas son distribuciones uniformes, lo que equivale a decir que saber una coordenada no me da ninguna información sobre la otra coordenada.

Ejemplo. A las 12 de la noche de un día de la semana comienzan a ser registrados las nuevas llamadas a un switch de telefonía. Sea X el instante de llegada de la primera llamada, medida en segundos transcurridos tras la medianoche. Sea Y el instante de llegada de la segunda llamada. En el modelo más

habitual utilizado en telefonía, X e Y son variables aleatorias continuas con densidad conjunta dada por

$$f_{X,Y}(x,y) = \begin{cases} \lambda^2 e^{-\lambda y} \text{ si } 0 \leq x < y \\ 0 \text{ en otro caso} \end{cases},$$

donde λ es una constante positiva. Vamos a calcular las distribuciones marginales y condicionadas que pueden darse:

- Marginal de X:
$$f_X(x) = \int_x^\infty \lambda^2 e^{-\lambda y} dy = \lambda e^{-\lambda x} \text{ si } 0 \leq x,$$
luego se trata de una distribución exponencial de parámetro λ.

- Marginal de Y :
$$f_Y(y) = \int_0^y \lambda^2 e^{-\lambda y} dx = \lambda^2 y e^{-\lambda y} \text{ si } y \geq 0.$$
Si nos fijamos, esta densidad es una $Gamma(2, \lambda)$, es decir una Erlang de parámetros 2 y λ.

- Condicionada de Y a los valores de X :
$$f_{Y/X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \lambda e^{-\lambda(y-x)} \text{ si } y > x.$$
En esta expresión no debe olvidarse que x es un valor fijo, dado.

- Condicionada de X a los valores de Y :
$$f_{X/Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{1}{y} \text{ si } 0 \leq x < y.$$
Es decir, conocido el instante en que llegó la segunda llamada (y), no se sabe nada de cuándo llegó la primera llamada, ya que la distribución de X condicionada a $Y = y$ es uniforme en $(0, y)$.

Ejemplo. Consideremos que la variable X representa el input de un canal de comunicación, con posibles valores $+1$ y -1 equiprobables, y sea Y el dígito que llega al destino, con valores también $+1$ y -1. El canal es un canal binario simétrico con probabilidad de cruce del 5 %.

Con los datos expuestos podemos caracterizar mediante sus funciones masa las distribuciones marginales de X e Y, la distribución conjunta de ambos y las dos distribuciones condicionadas posibles de cada variable respecto de la otra.

La distribución marginal de X viene dada por

$$f_X(x) = \begin{cases} \frac{1}{2} \text{ si } x = 1 \\ \frac{1}{2} \text{ si } x = -1 \end{cases}$$

La distribución marginal de Y viene dada por

$$\begin{aligned} P[Y=+1] &= P[Y=+1 \mid X=+1]\, P[X=+1] + P[Y=+1 \mid X=-1]\, P[X=-1] \\ &= 0.95 \times 0.5 + 0.05 \times 0.5 = 0.5 \\ P[Y=-1] &= 0.5, \end{aligned}$$

es decir

$$f_Y(y) = \begin{cases} \frac{1}{2} \text{ si } y = 1 \\ \frac{1}{2} \text{ si } y = -1 \end{cases}$$

La distribución de Y condicionada al suceso $X = +1$ viene dada por:

$$f_{Y|X=+1}(y) = \begin{cases} 0.95 \text{ si } y = 1 \\ 0.05 \text{ si } y = -1 \end{cases}$$

La distribución de Y condicionada al suceso $X = -1$ viene dada por:

$$f_{Y|X=-1}(y) = \begin{cases} 0.95 \text{ si } y = -1 \\ 0.05 \text{ si } y = 1 \end{cases}$$

La distribución conjunta de X e Y viene dada por

$$\begin{aligned} f_{X,Y}(x,y) &= P[Y=y \mid X=x]\, P[X=x] \\ &= \begin{cases} 0.95 \times 0.5 & si \quad x=+1, y=+1 \\ 0.05 \times 0.5 & si \quad x=+1, y=-1 \\ 0.05 \times 0.5 & si \quad x=-1, y=+1 \\ 0.95 \times 0.5 & si \quad x=-1, y=-1 \\ 0 & en\, otro\, caso \end{cases} \end{aligned}$$

La distribución de X condicionada al suceso $Y = +1$ viene dada por

$$f_{X|Y=+1}(x) = \frac{f_{X,Y}(x,+1)}{f_Y(+1)} = \begin{cases} 0.95 \text{ si } x = 1 \\ 0.05 \text{ si } x = -1 \end{cases}.$$

La distribución de X condicionada al suceso $Y = -1$ viene dada por

$$f_{X|Y=-1}(x) = \frac{f_{X,Y}(x,-1)}{f_Y(-1)} = \begin{cases} 0.05 \text{ si } x = 1 \\ 0.95 \text{ si } x = -1 \end{cases}.$$

5.3. Independencia estadística

En el capítulo referente a probabilidad hablamos de independencia de sucesos. Decíamos entonces que dos sucesos A y B eran independientes si y sólo si $P[A \cap B] = P[A] \cdot P[B]$.

Esta definición puede extenderse al caso en que tengamos dos variables aleatorias X e Y.

Concretamente, diremos que **X e Y son estadísticamente independientes** si y sólo si

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y),$$

donde $f_{X,Y}(\cdot)$, $f_X(\cdot)$ y $f_Y(\cdot)$ son función de densidad o función masa, dependiendo de si las variables son discretas o continuas.

La interpretación del hecho de que dos variables aleatorias sean estadísticamente independientes es que el comportamiento de una no tiene ningún efecto sobre la otra y viceversa. Cabe preguntarse en ese caso, qué sentido tiene una distribución condicionada de una variable a otra que no guarda ninguna relación con ella. Vamos a comprobarlo calculando las distribuciones condicionadas de variables aleatorias estadísticamente independientes:

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_X(x) \cdot f_Y(y)}{f_Y(y)} = f_X(x);$$

es decir, el comportamiento aleatorio de una variable aleatoria condicionada al valor de otra que es estadísticamente independiente de ella (descrito mediante la función $f_{X|Y=y}(x)$) es completamente igual que si no se condiciona a dicho valor (descrito por la función $f_X(x)$).

Ejemplo. Sea el vector (X,Y) con función de densidad conjunta

$$f_{X,Y}(x,y) = \begin{cases} 24xy \text{ si } x,y \geq 0 \text{ y } x+y \leq 1 \\ 0 \text{ en otro caso} \end{cases}.$$

La función de densidad marginal de X :

$$f_X(x) = \int_0^{1-x} 24xy \cdot dy = 12x(1-x)^2 \text{ si } 0 \leq x \leq 1$$

La función de densidad marginal de Y:

$$f_Y(y) = \int_0^{1-y} 24xy \cdot dx = 12y(1-y)^2 \text{ si } 0 \leq y \leq 1.$$

Como

$$f_{X,Y}(x,y) \neq f_X(x) \cdot f_Y(y),$$

las variables X e Y no son independientes.

Ejemplo. Sea ahora el vector (X,Y) con función de densidad conjunta

$$f_{X,Y}(x,y) = \begin{cases} 4xy \text{ si } 0 \leq x,y \text{ y } x,y \leq 1 \\ 0 \text{ en otro caso} \end{cases}$$

La función de densidad marginal de X:

$$f_X(x) = \int_0^1 4xy \cdot dy = 2x \text{ si } 0 \leq x \leq 1$$

La función de densidad marginal de Y:

$$f_Y(y) = \int_0^1 4xy \cdot dx = 2y \text{ si } 0 \leq y \leq 1.$$

Como

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y),$$

las variables aleatorias X e Y son independientes.

Ejemplo. Supongamos que dos componentes electrónicas tienen una duración cuya distribución de probabilidad puede considerarse exponencial de parámetro $\lambda = 2\ horas^{-1}$. Las componentes funcionan en paralelo, por lo que podemos considerar que son independientes. Por lo tanto, su función de densidad conjunta será

$$f_{X,Y}(x,y) = 2e^{-2x}2e^{-2y} = 4e^{-2(x+y)}$$

si $x, y > 0$.

¿Cuál será la probabilidad de que alguna de las componentes dure más de dos horas? Podemos plantearlo como

$$\begin{aligned} P[X > 2 \cup Y > 2] &= P[X > 2] + P[Y > 2] - P[X > 2 \cap Y > 2] \\ &= P[X > 2] + P[Y > 2] - P[X > 2] P[Y > 2], \end{aligned}$$

donde se ha utilizado en la probabilidad de la intersección el hecho de que las variables son independientes. Ahora sólo bastaría recordar que $P[X > 2] = e^{-2\times 2}$ y $P[Y > 2] = e^{-2\times 2}$.

¿Cuál sería la probabilidad de que la duración total de ambas componentes sea inferior a dos horas? La duración total vendría dada por $X + Y$, luego se nos pregunta por

$$\begin{aligned} P[X + Y < 2] &= \int_0^2 \int_0^{2-x} 4e^{-2(x+y)} dy dx \\ &= \int_0^2 \left[2e^{-2x}\left(1 - e^{-2(2-x)}\right)\right] dx \\ &= \int_0^2 \left(2e^{-2x} - 2e^{-4}\right) dx \\ &= \left(1 - e^{-4}\right) - 2e^{-4} \times 2 \\ &= 1 - 5e^{-4} \end{aligned}$$

De la interpretación que hemos dado de variables independientes se sigue de manera inmediata que si dos variables aleatorias son independientes, esto es, no mantienen ninguna relación, tampoco lo harán funciones

suyas. Este hecho se recoge en el siguiente resultado. Lo podemos enunciar más formalmente diciendo que si X e Y son variables aleatorias independientes y $V = g(X)$ y $W = h(Y)$ son funciones suyas, entonces, V y W también son independientes.

En el ámbito de las Telecomunicaciones se dan numerosas situaciones donde aparece una variable aleatoria W, suma de otras dos variables aleatorias (generalmente continuas) estadísticamente independientes, X e Y, es decir, $W = X + Y$. Por ejemplo, se da cuando a una señal X se le adhiere un ruido que le es completamente ajeno (independiente), Y. En ese caso, la suma representa la señal resultante y querremos conocer su comportamiento aleatorio a partir del de X e Y. Esto se conoce como **teorema de convolución**.

Concretamente, sean X e Y dos variables aleatorias independientes y sea $W = X + Y$. Entonces:
Si X e Y son continuas,

$$\begin{aligned} f_W(w) &= \int_{-\infty}^{\infty} f_Y(y) \cdot f_X(w-y) \cdot dy \\ &= f_X * f_Y(w) \end{aligned}$$

donde f_X y f_Y son las funciones de densidad de X e Y, respectivamente.
Si X e Y son discretas,

$$\begin{aligned} f_W(w) &= \sum_y f_Y(y) \cdot f_X(w-y) \\ &= f_X * f_Y(w) \end{aligned}$$

donde f_X y f_Y son las funciones masa de X e Y, respectivamente.

Ejemplo. Un sistema opera con una componente clave cuya duración, T_1, sigue una distribución exponencial de parámetro λ. Si esta componente falla, inmediatamente se pone en funcionamiento una componente exactamente igual que hasta entonces ha funcionado en *standby*, cuya duración notamos por T_2, variable aleatoria independiente de T_1.

Si pretendemos conocer la distribución de probabilidad de la duración total del sistema, que vendrá dada por la variable aleatoria $T = T_1 + T_2$, podemos poner en práctica el teorema de convolución. Para ello, tengamos en cuenta que

$$f_{T_i}(x) = \lambda e^{-\lambda x},\ i = 1, 2,$$

para $x > 0$. Por tanto,

$$f_T(z) = \int_0^z \lambda e^{-\lambda x} \lambda e^{-\lambda(z-x)} dx = \lambda^2 z e^{-\lambda z}$$

para $z > 0$. Como vemos, se trata de una distribución Erlang de parámetros 2 y λ. Si recordamos, esta era una de las caracterizaciones de la distribución Erlang, suma de exponenciales independientes.

En el caso de que en vez de dos variables aleatorias se tenga un vector $X = (X_1, ..., X_N)'$, la manera natural de extender el concepto de independencia es inmediata.

Se dice que el vector está formado por **componentes independientes** si

$$f_{X_1,...,X_N}(x_1,...,x_N) = f_{X_1}(x_1) \cdot ... \cdot f_{X_N}(x_N).$$

Finalmente, si se tienen dos vectores aleatorios $X_{N\times 1}$ e $Y_{M\times 1}$, se dice que son **independientes** si

$$f_{X,Y}(x_1,...,x_N,y_1,...,y_M) = f_X(x_1,...,x_N)\, f_Y(y_1,...,y_M).$$

5.4. Medias, varianzas y covarianzas asociadas a un vector aleatorio

Si tenemos un vector aleatorio formado por las variables aleatorias $X_1,...,X_N$ y $g(\cdot)$ es una función de estas variables, entonces, la **media o esperanza matemática** de esta función es

$$E[g(X_1,...,X_N)] = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} g(x_1,...,x_N) \cdot f_{X_1,...,X_N}(x_1,...,x_N) \cdot dx_N \cdot ... \cdot dx_1$$

donde $f_{X_1,...,X_N}(x_1,...,x_N)$ es la función de densidad o la función masa del vector aleatorio (entendiendo en este último caso la integral como una suma).

Como consecuencia inmediata de esta definición, tenemos una primera e importante propiedad: este operador esperanza multivariante también es lineal, en el sentido que se recoge en el siguiente resultado.

Concretamente, podemos formalizarlo diciendo que si tenemos un vector aleatorio $(X_1,...,X_N)'$ y $\alpha_1,...,\alpha_N$ escalares cualesquiera, entonces

$$E[\alpha_1 X_1 + ... + \alpha_N X_N] = \alpha_1 E[X_1] + ... + \alpha_N E[X_N],$$

es decir, la media de la suma ponderada es la suma ponderada de las medias. Podemos tratar de recordar este resultado si pensamos que es exactamente la misma propiedad que tiene el operador integral, que *parte las sumas y saca fuera los escalares.*

5.4.1. Covarianza y coeficiente de correlación lineal

Anteriormente hemos comentado que estudiar vectores aleatorios desde una perspectiva estadística tiene sentido, sobre todo, porque permite analizar las relaciones que se dan entre las variables del vector. Por ejemplo, vimos cómo los valores de una variable pueden afectar en mayor o menor medida a la distribución de probabilidad de las otras variables.

Sin embargo, sería muy interesante disponer de una medida numérica sencilla de calcular y de interpretar para cuantificar al menos en parte cuál es el grado de relación existente entre dos variables de un vector aleatorio.

En este sentido, dado el vector aleatorio (X, Y), se define la **correlación entre** X e Y como

$$R_{XY} = m_{11} = E[XY],$$

a partir de la cual se puede calcular la **covarianza entre** X e Y como

$$Cov(X, Y) = E[(X - EX) \cdot (Y - EY)] = E[XY] - EX \cdot EY = R_{XY} - EX \cdot EY.$$

La covarianza entre dos variables[2] es una medida de la asociación lineal existente entre ellas. Será positiva si la relación entre ambas es directa (si crece una crece la otra) y negativa si es inversa (si crece una decrece la otra); además, será tanto mayor en valor absoluto cuanto más fuerte sea la relación lineal existente.

Para poder valorar esta relación lineal en términos relativos se estandariza la covarianza, dando lugar a lo que se conoce como **coeficiente de correlación lineal**:

$$\rho = \frac{Cov[X, Y]}{\sqrt{Var[X] \cdot Var[Y]}}.$$

Vamos a detallar claramente los posibles valores de ρ y su interpretación:

- Este coeficiente es siempre un número real entre -1 y 1.
- Si es cero, indica una ausencia total de relación lineal entre las variables.
- Si es uno o menos uno indica una relación lineal total entre las variables, directa o inversa según lo indique el signo (esto lo veremos enseguida).
- En la medida en que esté más lejos del cero indica una relación lineal más intensa entre las variables.

Si dos variables aleatorias tienen covarianza cero o equivalentemente, si $R_{XY} = EX \cdot EY$, se dicen que son **incorreladas**. Por su parte, si dos variables aleatorias son tales que $R_{XY} = 0$, se dice que son **ortogonales**.

Dos variables aleatorias son incorreladas si carecen de cualquier tipo de relación lineal. Por otra parte, definimos anteriormente el concepto de independencia entre variable aleatoria, que implicaba la ausencia de relación entre ellas. Tenemos, así, dos conceptos, independencia e incorrelación, que están bastante relacionados.

En concreto, dos variable aleatoria independientes, X e Y, son siempre incorreladas, es decir, $\rho_{X,Y} = 0$. La razón es que, por ser independientes,

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y),$$

[2] Si se considera la covarianza de una variable aleatoria consigo misma,

$$Cov(X, X) = E[(X - EX)(X - EX)] = E\left[(X - EX)^2\right] = VarX,$$

esta cantidad coincide con su varianza.

luego

$$\begin{aligned} R_{XY} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\cdot f_X(x)\cdot f_Y(y)\cdot dy\cdot dx \\ &= \int_{-\infty}^{\infty} x f_X(x)\,dx\cdot\int_{-\infty}^{\infty} y f_Y(y)\,dy = EX\cdot EY, \end{aligned}$$

en cuyo caso $Cov[X,Y]=0$.

La pregunta obvia que surge a la luz de este resultado es: ¿ y al contrario? ¿Dos variable aleatoria incorreladas serán independientes? O equivalentemente, ¿si dos variable aleatoria no tienen ninguna relación de tipo lineal (incorreladas), ocurrirá que tampoco tienen ninguna relación de ningún tipo (independientes)? La respuesta es que no en general.

Ejemplo. Sea α una variable aleatoria con distribución uniforme en $(0,2\pi)$. Sean

$$\begin{aligned} X &= \cos\alpha \\ Y &= \sin\alpha. \end{aligned}$$

Se tiene que

$$\begin{aligned} EX &= \int_0^{2\pi}\cos\alpha\frac{1}{2\pi}d\alpha = 0 \\ EY &= \int_0^{2\pi}\sin\alpha\frac{1}{2\pi}d\alpha = 0 \\ E[XY] &= \int_0^{2\pi}\sin\alpha\cos\alpha\frac{1}{2\pi}d\alpha \\ &= \frac{1}{2\pi}\int_0^{2\pi}\sin 2\alpha d\alpha = 0, \end{aligned}$$

por lo que X e Y son variables incorreladas. Sin embargo, puede demostrarse fácilmente que no son independientes.

Nota. La relación más fuerte de tipo lineal que puede darse corresponde al caso en que una variable aleatoria Y es exactamente una combinación lineal de otra, X, es decir, $Y=aX+b$. En ese caso,

$$\rho_{XY} = 1\cdot signo(a).$$

La demostración es muy sencilla. Tengamos en cuenta que

$$E[XY] = E[X(aX+b)] = aE\left[X^2\right] + bE[X],$$

luego

$$\begin{aligned} Cov(X,Y) &= E[XY] - EX \cdot EY \\ &= aE\left[X^2\right] + bE[X] - EX(aEX+b) \\ &= a\left(E\left[X^2\right] - EX^2\right) = aVarX \\ VarY &= E\left[((aX+b)-(aEX+b))^2\right] \\ &= E\left[(aX-aEX)^2\right] = E\left[a^2(X-EX)^2\right] \\ &= a^2 E\left[(X-EX)^2\right] = a^2 VarX, \end{aligned}$$

y

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{VarX \cdot VarY}} = \frac{aVarX}{\sqrt{VarXa^2VarX}} = 1 \cdot signo(a).$$

Nota. Es importante insistir en que la covarianza y su versión estandarizada, el coeficiente de correlación lineal, proporcionan una medida de la relación **lineal**, no de otro tipo. Por ejemplo, supongamos que la Figura 5.3 representa los valores conjuntos de dos variables X e Y. Está claro que ambas guardan una clarísima relación dada por una parábola: de hecho, $Y = X^2$. Sin embargo, el coeficiente de correlación lineal entre ambas será muy bajo, ya que en realidad, la relación que las une no es lineal en absoluto, sino parabólica. En este caso, lo recomendable sería, a la vista del gráfico, decir que sí existe una fuerte relación lineal entre X e $\pm\sqrt{Y}$.

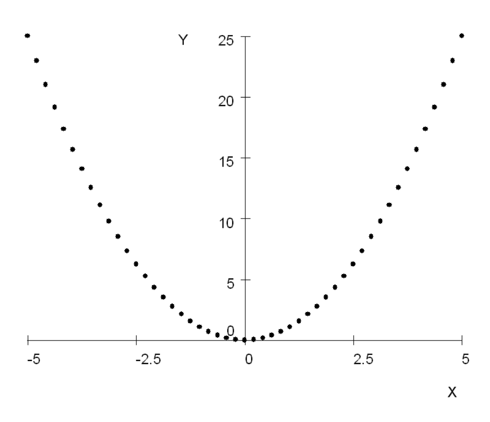


Figura 5.3: Muestra conjunta de valores de dos variables aleatorias.

Cuando se tienen muestras de pares de variables aleatorias, podemos calcular la versión muestral del coeficiente de correlación lineal. Esa versión muestral dará una estimación del verdadero valor del coeficiente de correlación (poblacional). Esta cuestión se aborda con más detalle en el capítulo de regresión. Aquí tan sólo queremos plasmar con ejemplos cómo se traduce el hecho de que dos variables tengan un mayor o menor coeficiente de correlación. En la Figura 5.4 observamos representaciones conjuntas de muestras de pares de variables en unos ejes cartesianos (nubes de puntos). Cada punto de cada eje cartesiano representa un valor

dado de la muestra del par (X, Y). Aparecen 4 figuras, correspondientes a 4 simulaciones de pares de variables (X, Y) con distintos coeficientes de correlación.

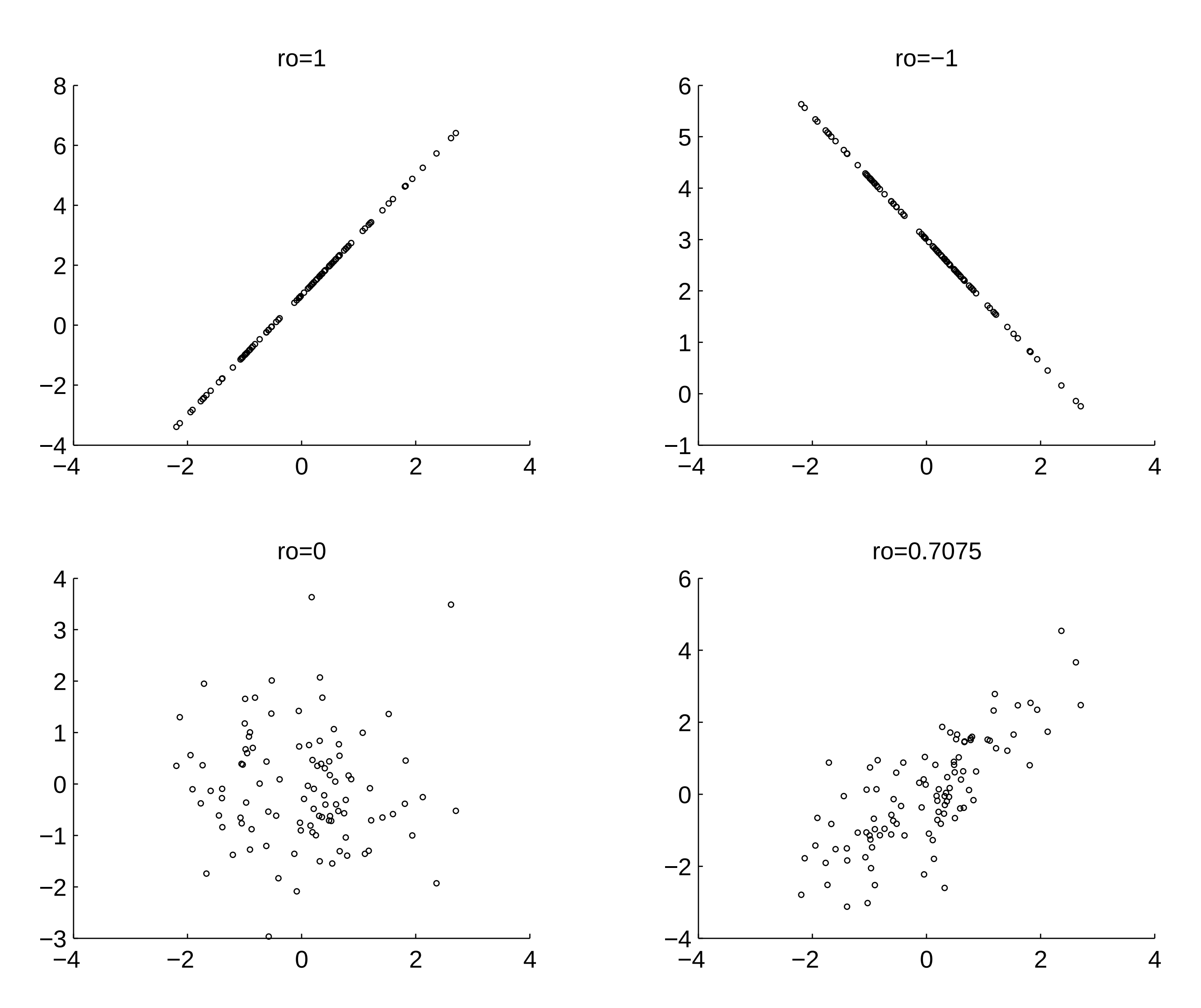


Figura 5.4: Nubes de puntos correspondientes a distintos posibles coeficientes de correlación lineal.

Ejemplo. Sean X e Y las variable aleatoria que miden el tiempo que transcurre hasta la primera y la segunda llamada, respectivamente, a una centralita telefónica. La densidad conjunta de estas variables es $f_{X,Y}(x, y) = e^{-y}$ para $0 < x < y$. En un ejemplo anterior ya vimos que, lógicamente, el tiempo hasta la segunda llamada depende del tiempo hasta la primera llamada, pero ¿en qué grado? Vamos a abordar este problema calculando el coeficiente de correlación lineal entre ambas variables.

Como $\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{VarXVarY}}$, tenemos que calcular $Cov(X,Y)$, $VarX$ y $VarY$.

$$\begin{aligned} E[XY] &= \int\int xy f_{X,Y}(x,y)\,dxdy \\ &= \int_0^\infty \int_0^y xye^{-y}dxdy = \int_0^\infty ye^{-y}\left[\frac{x^2}{2}\right]_0^y dy \\ &= \int_0^\infty \frac{y^3}{2}e^{-y}dy = 3. \end{aligned}$$

$$f_X(x) = \int f_{X,Y}(x,y)\,dy = \int_x^\infty e^{-y}dy = e^{-x}, \text{ para } x > 0,$$

luego

$$EX = \int x f_X(x)\,dx = \int_0^\infty xe^{-x}dx = 1.$$

$$f_Y(y) = \int f_{X,Y}(x,y)\,dx = \int_0^y e^{-y}dx = ye^{-y}, \text{ para } y > 0,$$

luego

$$EY = \int y f_Y(y)\,dy = \int_0^\infty y^2 e^{-y}dy = 2.$$

Por tanto,

$$Cov(X,Y) = 3 - 1 \times 2 = 1.$$

Por su parte,

$$\begin{aligned} E\left[X^2\right] &= \int x^2 f_X(x)\,dx = \int_0^\infty x^2 e^{-x}dx = 2 \\ VarX &= 2 - 1^2 = 1 \end{aligned}$$

y

$$\begin{aligned} E\left[Y^2\right] &= \int y^2 f_Y(y)\,dy = \int_0^\infty y^3 e^{-y}dy = 6 \\ VarY &= 6 - 2^2 = 2, \end{aligned}$$

así que, finalmente,

$$\rho_{X,Y} = \frac{1}{\sqrt{1 \times 2}} = 0.707.$$

El resultado indica que, en efecto, el grado de relación lineal es alto y directo.

Las propiedades del operador esperanza son muy útiles en la práctica, por ejemplo, cuando se trata de conocer la varianza de combinaciones lineales de varias variables. Veamos algún ejemplo al respecto y después un resultado general que los englobe todos.

Ejemplo. Calculemos la varianza de $X_1 + X_2$:

$$E\left[(X_1 + X_2)^2\right] = E\left[X_1^2 + X_2^2 + 2X_1X_2\right] = E\left[X_1^2\right] + E\left[X_2^2\right] + 2E\left[X_1X_2\right]$$

$$\begin{aligned} Var\left(X_1 + X_2\right) &= E\left[(X_1 + X_2)^2\right] - E\left[X_1 + X_2\right]^2 \\ &= E\left[X_1^2\right] + E\left[X_2^2\right] + 2E\left[X_1X_2\right] - \left(EX_1 + EX_2\right)^2 \\ &= E\left[X_1^2\right] + E\left[X_2^2\right] + 2E\left[X_1X_2\right] - EX_1^2 - EX_2^2 - 2EX_1EX_2 \\ &= E\left[X_1^2\right] - EX_1^2 + E\left[X_2^2\right] - EX_2^2 + 2\left(E\left[X_1X_2 - EX_1EX_2\right]\right) \\ &= VarX_1 + VarX_2 + 2Cov\left(X_1, X_2\right). \end{aligned}$$

Ejemplo. Calculemos la varianza de $X_1 - X_2$:

$$E\left[(X_1 - X_2)^2\right] = E\left[X_1^2 + X_2^2 - 2X_1X_2\right] = E\left[X_1^2\right] + E\left[X_2^2\right] - 2E\left[X_1X_2\right]$$

$$\begin{aligned} Var\left(X_1 - X_2\right) &= E\left[(X_1 - X_2)^2\right] - E\left[X_1 - X_2\right]^2 \\ &= E\left[X_1^2\right] + E\left[X_2^2\right] - 2E\left[X_1X_2\right] - \left(EX_1 - EX_2\right)^2 \\ &= E\left[X_1^2\right] + E\left[X_2^2\right] - 2E\left[X_1X_2\right] - EX_1^2 - EX_2^2 + 2EX_1EX_2 \\ &= E\left[X_1^2\right] - EX_1^2 + E\left[X_2^2\right] - EX_2^2 - 2\left(E\left[X_1X_2 - EX_1EX_2\right]\right) \\ &= VarX_1 + VarX_2 - 2Cov\left(X_1, X_2\right). \end{aligned}$$

Podemos generalizar estos ejemplos en el siguiente resultado. Sea una suma de $N-$variables, $X = \sum_{i=1}^{N} \alpha_i \cdot X_i$. Entonces,

$$Var\left[X\right] = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \cdot \alpha_j \cdot Cov\left(X_i, X_j\right),$$

donde $Cov\left(X_i, X_i\right) = Var\left(X_i\right)$, para $i = 1, ..., N$.

La demostración es bien sencilla. Como $\bar{X} = \sum_{i=1}^{N} \alpha_i \cdot EX_i$,

$$\begin{aligned} Var[X] &= E\left[\left(X - \bar{X}\right)^2\right] \\ &= E\left[\left(\sum_{i=1}^{N} \alpha_i \cdot \left(X_i - \bar{X}_i\right)\right)\left(\sum_{i=1}^{N} \alpha_i \cdot \left(X_i - \bar{X}_i\right)\right)\right] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \cdot \alpha_j \cdot E\left[\left(X_i - \bar{X}_i\right)\left(X_j - \bar{X}_j\right)\right] \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \cdot \alpha_j \cdot Cov\left(X_i, X_j\right) \end{aligned}$$

Fijémonos que, en el caso en que las variables sean incorreladas,

$$Var[X] = \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \cdot \alpha_j \cdot Cov\left(X_i, X_j\right) = \sum_{i=1}^{N} \alpha_i^2 \cdot Var[X_i],$$

ya que

$$Cov[X, Y] = \begin{cases} 0 \text{ si } i \neq j \\ Var[X_i] \text{ si } i = j \end{cases}.$$

5.4.2. Vector de medias y matriz de varianzas-covarianzas de un vector

Dado un vector de $N-$variables, $X = (X_1, ..., X_N)'$, se define su **vector de medias** como

$$\mu_X = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_N] \end{pmatrix},$$

y su **matriz de varianzas-covarianzas** como

$$C_X = \left(C_{i,j}\right)_{i,j=1,\ldots,N},$$

donde

$$C_{i,j} = \begin{cases} Var\left(X_i\right) \text{ si } i = j \\ Cov\left(X_i, X_j\right) \text{ si } i \neq j \end{cases}.$$

Esta matriz contiene las varianzas de cada variable del vector en la diagonal y en el elemento (i, j) la covarianza entre la $i-$ésima y la $j-$ésima variable.

En forma matricial, la matriz de covarianzas puede definirse como

$$C_{X\ N\times N} = E\left[\left(X - \mu_X\right)_{N\times 1}\left(X - \mu_X\right)'_{1\times N}\right].$$

Por otra parte,

$$C_X = E\left[\left(X - \mu_X\right)\left(X - \mu_X\right)'\right] = E\left[XX'\right] - \mu_X \mu'_X,$$

donde a la matriz $E[XX']$ se le suele denominar **matriz de correlaciones o de autocorrelaciones**, y se le nota R_X.

Ambas matrices, C_X y R_X, son matrices simétricas.

La linealidad del operador media facilita rápidamente la expresión del vector de medias y la matriz de varianzas-covarianzas de combinaciones lineales de vectores, como se recoge en el siguiente resultado. Concretamente, si tenemos el vector aleatorio $X_{N\times 1}$ con vector de medias μ_X y matriz de varianzas covarianzas C_X y el vector $Y_{M\times 1} = A_{M\times N} \cdot X_{N\times 1} + b_{M\times 1}$, entonces, el vector de medias y la matriz de varianzas covarianzas de Y vienen dadas por

$$\mu_Y = A\mu_X + b$$
$$C_Y = AC_X A'.$$

Ejemplo. Vamos a ver que la aplicación de este resultado facilita bastante determinados cálculos. Por ejemplo, si queremos calcular $Var(X_1 + X_2)$, podemos tener en cuenta que

$$X_1 + X_2 = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},$$

de manera que

$$Var(X_1 + X_2) = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} VarX_1 & Cov(X_1, X_2) \\ Cov(X_1, X_2) & VarX_2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$
$$= VarX_1 + VarX_2 + 2Cov(X_1, X_2).$$

De igual forma, si queremos calcular $Var(5X_1 - 3X_2)$, dado que

$$5X_1 - 3X_2 = \begin{pmatrix} 5 & -3 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},$$

se tiene que

$$Var(5X_1 - 3X_2) = \begin{pmatrix} 5 & -3 \end{pmatrix} \begin{pmatrix} VarX_1 & Cov(X_1, X_2) \\ Cov(X_1, X_2) & VarX_2 \end{pmatrix} \begin{pmatrix} 5 \\ -3 \end{pmatrix}$$
$$= 25VarX_1 + 9VarX_2 - 30Cov(X_1, X_2).$$

5.5. Distribución normal multivariante

En el contexto de los modelos de distribuciones de probabilidad para variables aleatorias, la distribución normal constituye el ejemplo más relevante, tanto por la frecuencia de su aplicación en casos reales como por la gran versatilidad de sus propiedades matemática. En el contexto de los vectores aleatorios que estamos tratando en este capítulo, nos ocupamos de la versión multivariante de esta distribución. De nuevo podemos

estar seguros de que se trata del caso más interesante por dos motivos: porque aparece como modelo adecuado en un gran número de fenómenos de la naturaleza y porque sus propiedades matemáticas on inmejorables.

Un vector formado por N variables aleatorias $X = (X_1, ..., X_N)'$ se dice que sigue una **distribución normal multivariante o distribución conjuntamente normal o conjuntamente gaussiana,** con vector de medias μ_X y matriz de varianzas-covarianzas C_X, si su función de densidad conjunta es de la forma

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^N \det(C_X)}} \cdot \exp\left[-\frac{1}{2}(x-\mu_X)' \cdot C_X^{-1}(x-\mu_x)\right],$$

donde

$$\begin{aligned} C_X &= (C_{i,j})_{i,j=1,...,N} \\ C_{ij} &= \begin{cases} Var[X_i] & \text{si } i = j \\ Cov[X_i, X_j] & \text{si } i \neq j \end{cases} \\ x &= (x_1, ..., x_N)' \\ \mu_X &= (EX_1, ..., EX_N)' \end{aligned}$$

y se nota $X \to N_N(\mu_X; C_X)$.

Vamos a destacar algunas de las excelentes propiedades de la distribución normal multivariante. Concretamente, nos centraremos en los siguientes resultados:

- Cualquier marginal sigue también una distribución normal.
- Cualquier distribución condicionada sigue también una distribución normal.
- Cualquier combinación lineal de un vector normal es también normal.

Vamos a concretarlos. En primer lugar, si tenemos un vector $X_{N\times 1} = (X_1, ..., X_N)'$ con distribución conjuntamente gaussiana de vector de medias μ y matriz de covarianzas C_X, en ese caso, el subconjunto de variables del vector, $(X_{i1}, ..., X_{iM})$, con $M < N$ también sigue distribución conjuntamente gaussiana, de parámetros $(\mu_{i1}, ..., \mu_{iM})'$ y matriz de covarianzas constituida por las filas y las columnas de C_X correspondientes a las variables $X_{i1}, ..., X_{iM}$.

Ejemplo. Sea un vector $(X_1, X_2, X_3)'$ gaussiano, de vector de medias cero y matriz de covarianzas

$$\begin{pmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

En aplicación del resultado anterior, las marginales univariantes siguen las distribuciones siguientes: $X_1 \to N(0,2)$, $X_2 \to N(0,3)$, $X_3 \to N(0,1)$.

Por su parte, las marginales bivariantes siguen las distribuciones siguientes:

$$(X_1, X_2)' \to N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}\right)$$
$$(X_1, X_3)' \to N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}\right)$$
$$(X_2, X_3)' \to N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}\right)$$

En cuanto a las distribuciones condicionales, cualquier subconjunto de variables de un vector gaussiano condicionado a los valores de cualquier otro subconjunto de variables del propio vector sigue distribución conjuntamente gaussiana. Concretamente, la distribución de $X_{N\times 1}$ condicionada a $Y_{M\times 1} = y_{M\times 1}$, siendo $(X, Y)'_{(M+N)\times 1}$ conjuntamente gaussiano, es gaussiana de vector de medias

$$E\left[\mathbf{X}\,|_{\mathbf{Y}=\mathbf{y}}\right] = \mu^{\mathbf{X}}_{N\times 1} + (C_{\mathbf{XY}})_{N\times M}\left(C_Y^{-1}\right)_{M\times M}\left(y_{M\times 1} - \mu^{\mathbf{Y}}_{M\times 1}\right)$$

y matriz de varianzas-covarianzas

$$Var\left(\mathbf{X}\,|_{\mathbf{Y}=y}\right) = C_X - C_{XY}C_Y^{-1}C'_{XY},$$

donde el elemento (i, j) de C_{XY} es $Cov\left(X_i, Y_j\right)$.

Ejemplo. Siguiendo con el ejemplo anterior, vamos a considerar la distribución de X_1 condicionada a $(X_2, X_3)' = (0.5, 0.25)'$.

Según el resultado, ésta es gaussiana, de vector de medias

$$E\left[X_1\,|_{X_2=0.5,\ X_3=0.25}\right] = 0 + \begin{pmatrix} 1 & 0 \end{pmatrix}\begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}^{-1}\begin{pmatrix} 0.5-0 \\ 0.25-0 \end{pmatrix} = 0.125$$

y matriz de covarianzas (es decir, varianza)

$$Var\left(X_1\,|_{X_2=0.5,\ X_3=0.25}\right) = 2 - \begin{pmatrix} 1 & 0 \end{pmatrix}\begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}^{-1}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1.5$$

Ejemplo. Como caso particular, vamos a describir con más detalle el caso bivariante, tanto en lo que respecta a su densidad como a las distribuciones marginales y condicionadas.

Sea por tanto un vector $(X, Y)'_{2\times 1}$, con distribución conjuntamente gaussiana de vector de medias

$(\mu_X, \mu_Y)'$ y matriz de covarianzas

$$C_{(X,Y)} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix},$$

donde $\rho = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$ es el coeficiente de correlación lineal. Entonces, $\det C_{(X,Y)} = \sigma_X^2\sigma_Y^2\left(1-\rho^2\right)$ y

$$C_{(X,Y)}^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_X^2} & -\frac{\rho}{\sigma_X\sigma_Y} \\ -\frac{\rho}{\sigma_X\sigma_Y} & \frac{1}{\sigma_Y^2} \end{pmatrix}.$$

Por tanto, la función de densidad conjunta es

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \\ &\cdot \exp\left\{\frac{-1}{2\left(1-\rho^2\right)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho\,(x-\mu_x)\,(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}. \end{aligned}$$

Esta función alcanza su máximo, $\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$, en el punto (μ_X, μ_Y).

Evidentemente, las distribuciones marginales son $N\left(\mu_X, \sigma_X^2\right)$ y $N\left(\mu_Y, \sigma_Y^2\right)$.

En lo que respecta a las distribuciones condicionadas, aplicando el último resultado tenemos que

$$\begin{aligned} X \mid Y = y_0 &\to N\left(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y_0 - \mu_Y); \sigma_X^2\left(1-\rho^2\right)\right) \\ Y \mid X = x_0 &\to N\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x_0 - \mu_X); \sigma_Y^2\left(1-\rho^2\right)\right). \end{aligned}$$

Obsérvese que, curiosamente, la varianza condicionada no depende del valor que condiciona. Esto tendrá importantes repercusiones más adelante.

Continuando con las propiedades, una de las más útiles es su invarianza frente a transformaciones lineales. Concretamente, si tenemos un vector aleatorio $\mathbf{X}_{N\times 1} = (X_1, ..., X_N)'$ con distribución gaussiana, vector de medias μ_X y matriz de covarianzas C_X, entonces una combinación lineal suya,

$$\mathbf{Y}_{M\times 1} = A_{M\times N} \cdot X_{N\times 1} + b_{M\times 1}$$

tiene distribución gaussiana de vector de medias $\mu_Y = A \cdot \mu_X + b$ y matriz de covarianzas $C_Y = A \cdot C_X \cdot A'$.

Ejemplo. Sean dos variable aleatoria X_1 y X_2 con distribución conjuntamente gaussiana con medias cero, varianzas $\sigma_{X_1}^2 = 4$ y $\sigma_{X_2}^2 = 9$ y covarianza, $c_{X_1,X_2} = 3$. Si estas variables se transforman linealmente

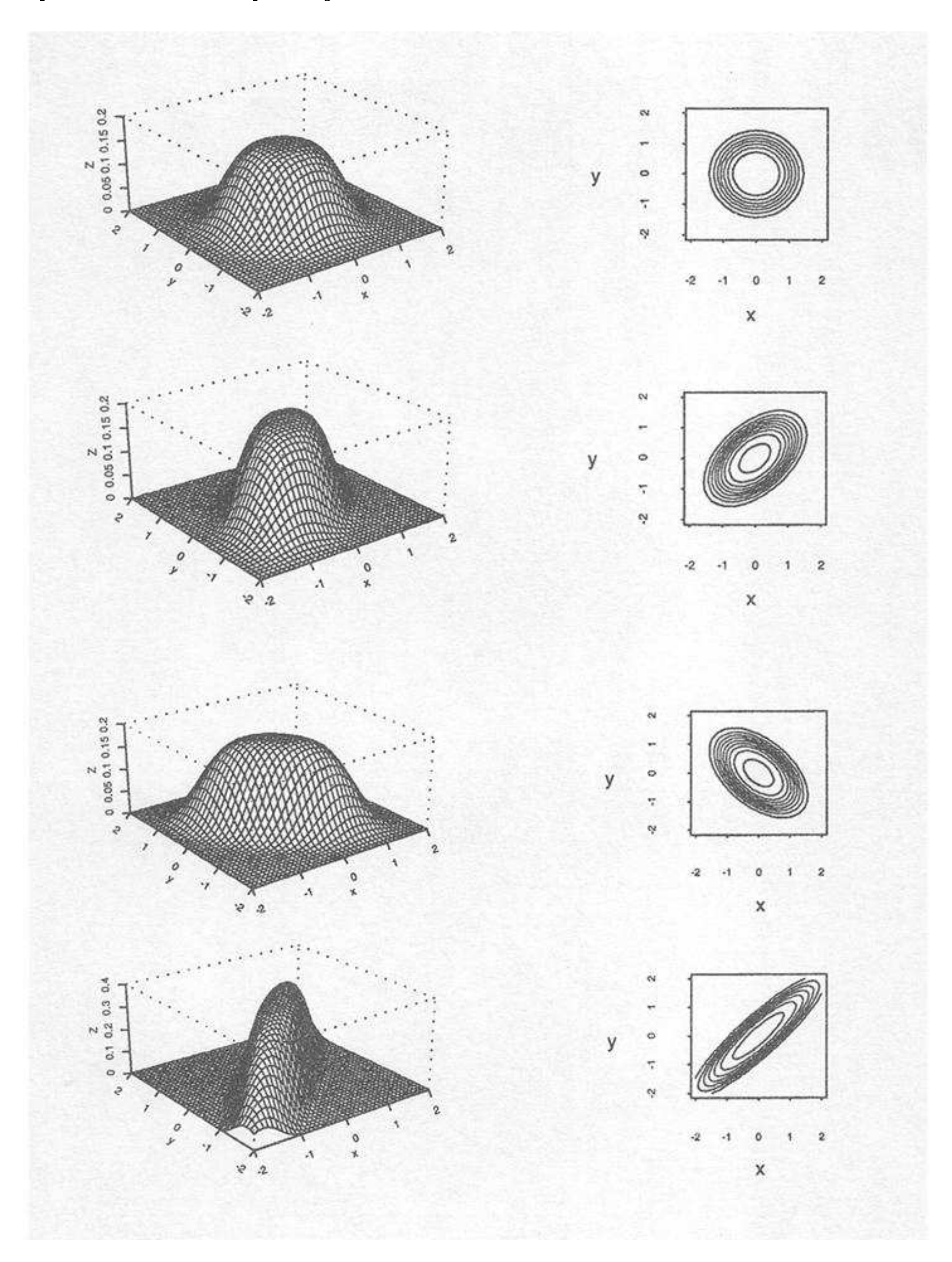


Figura 5.5: Ejemplos de densidades de la normal bivariantes con $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$ y $\rho = 0,\ 0.5,\ -0.5$ y 0.9. (En http://www.ilri.org/InfoServ/Webpub/Fulldocs/Linear_Mixed_Models/AppendixD.htm).

en las variables

$$Y_1 = X_1 - 2X_2$$
$$Y_2 = 3X_1 + 4X_2$$

las nuevas variables tienen distribución conjuntamente gaussiana, con medias

$$(\mu_{Y_1}, \mu_{Y_2})' = \begin{pmatrix} 1 & -2 \\ 3 & 4 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

y matriz de covarianzas

$$\begin{pmatrix} \sigma_{Y_1}^2 & c_{Y_1,Y_2} \\ c_{Y_1,Y_2} & \sigma_{Y_2}^2 \end{pmatrix} = \begin{pmatrix} 1 & -2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 4 & 3 \\ 3 & 9 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ -2 & 4 \end{pmatrix} = \begin{pmatrix} 28 & -66 \\ -66 & 252 \end{pmatrix}$$

Otra de las más importantes propiedades es que se trata del único caso en el que independencia e incorrelación son equivalentes. Es decir, si $X_{N\times 1}$ es un vector con distribución conjuntamente gaussiana, entonces sus componentes son incorreladas si y sólo si son independientes.

La demostración es sencilla. Ya sabemos que si son independientes son incorreladas (incluso si la distribución no es conjuntamente gaussiana). Por su parte, para probar que si son incorreladas entonces son independientes sólo hay que tener en cuenta que si son incorreladas, la matriz de covarianzas es diagonal y la densidad conjunta puede expresarse como producto de las marginales, ya que

$$\begin{aligned} f_X(x_1, ..., x_N) &= \frac{1}{\sqrt{(2\pi)^N \det(C_X)}} \exp\left\{-\frac{1}{2}(x-\mu_X)' C_X^{-1}(x-\mu_X)\right\} \\ &= \frac{1}{\sqrt{(2\pi)^N \sigma_1^2 ... \sigma_N^2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^N \left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right\} \\ &= \prod_{i=1}^N f_{X_i}(x_i). \end{aligned}$$

donde $x = (x_1, ..., x_N)'$, $\mu_X = (\mu_1, ..., \mu_N)'$ y

$$C_X = \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_N^2 \end{pmatrix}.$$

Parte III

Inferencia estadística

Capítulo 6

Distribuciones en el muestreo

> Pocas observaciones y mucho razonamiento conducen al error; muchas observaciones y poco razonamiento, a la verdad.
>
> Alexis Carrel

Resumen. En este capítulo se pretende llamar la atención acerca de que los parámetros muestrales son en realidad variables aleatorias. Se analiza así la distribución de probabilidad de la media muestral y de la varianza muestral en diversas situaciones.

Palabras clave: distribuciones en el muestreo, t de Student, F de Snedecor.

6.1. Introducción

Al estudiar el concepto de variable aleatoria, dijimos que viene motivado porque muchas de las variables que se observan en la vida real, en el ambiente de las Ingenierías en particular, están sujetas a incertidumbre.

Eso quiere decir que si nosotros obtenemos algunas observaciones de esas variables (muestras), los datos no son iguales. Es más, si obtenemos otras observaciones, las dos muestras tampoco serán ni mucho menos idénticas.

Por tanto, al hablar de distribuciones teóricas de probabilidad, lo que pretendíamos era proponer un modelo que permitiera calcular probabilidades asociadas, no a una muestra en particular de datos, sino a todas las posibles muestras, con todos los posibles datos de la variable.

Recordemos el ejemplo que pusimos: las distribuciones de probabilidad son como un traje que elegimos para ponernos cualquier día durante un periodo de tiempo amplio. En la medida que el traje de una variable, su distribución, *le quede bien*, los resultados que obtengamos mediante el cálculo de probabilidades podrán aplicarse a cualquier dato o conjunto de datos de la variable. Pero igualmente, si un traje (una distribución de probabilidad teórica) *no le queda bien* a una variable, los resultados teóricos, obtenidos a partir de una función masa o una función de densidad teóricas, pueden no ser realistas respecto a los resultados empíricos que se obtengan mediante muestras de la variable.

¿Qué nos queda por hacer a lo largo del curso? Dado que, en general, las distribuciones teóricas de probabilidad dependen de uno o más parámetros, lo que nos ocupará gran parte del resto del curso es tratar de elegir

adecuadamente esos parámetros. En el ejemplo de los trajes podíamos pensar que esto es como aprender a escoger la talla del traje.

En este capítulo vamos a comenzar con algunas cuestiones teóricas acerca de lo que implica el proceso de muestreo, previo a la elección de los parámetros y, posteriormente, nos vamos a centrar en resultados que implica el muestreo de datos de variables que siguen una distribución normal.

6.2. Muestreo aleatorio

En multitud de ámbitos de la vida real es evidente que la mejor forma de aprender algo es a partir de la experiencia. Eso quiere decir que solemos utilizar aquello que vemos para aprender pautas y conductas que luego generalizamos.

En Estadística pasa algo muy similar: necesitamos basarnos en muestras de una variable para poder aprender de ellas y generalizar, inferir, aspectos referentes a las muestras a toda la población.

Sin embargo, como en la vida real, en Estadística también debemos ser muy cuidadosos con los datos sobre los que basamos nuestro aprendizaje. ¿Qué pasaría si basamos nuestro aprendizaje en experiencias incorrectas o poco significativas?

Para que esto no ocurra debemos basarnos en muestras donde todos los individuos de la población puedan verse representados. Por otra parte, es evidente que cuanto mayores sean las muestras más fiables deberían ser nuestras inferencias.

El concepto clave en este planteamiento es el de *muestra aleatoria simple*. Supongamos que estamos observando una variable aleatoria, X, en una población determinada. Ya dijimos que una muestra aleatoria simple de X consiste en la recopilación de datos de la variable, mediante la repetición del experimento al que está asociada, con dos condiciones básicas:

1. Que todos los elementos de la población tengan las mismas posibilidades de salir en la muestra.

2. Que las distintas observaciones de la muestra sean independientes entre sí.

En ese caso, los valores que toma la variable en cada una de las observaciones de una muestra de tamaño n, $X_1, ..., X_n$, son en sí mismos, variables aleatorias independientes que siguen la misma distribución de probabilidad, llamada **distribución poblacional**. Esta distribución es, en principio, desconocida, por lo que se intentará utilizar la muestra para hacer inferencia sobre ella y, al menos, aproximar la forma de esta distribución.

6.3. Distribuciones en el muestreo

Supongamos que estamos observando una variable aleatoria X, y que obtenemos una muestra aleatoria simple suya, $x_1^1, ..., x_n^1$. Con esos datos podemos calcular la media de la muestra, $\bar{x}_1$, y la desviación típica de la muestra, s_1, por ejemplo.

Pero debemos ser conscientes de lo que significa muestra *aleatoria*. El hecho de que hayan salido los valores $x_1^1, ..., x_n^1$ es fruto del azar. De hecho, si obtenemos otra muestra, $x_1^2, ..., x_n^2$, obtendremos otra media, $\bar{x}_2$ y otra desviación típica de la muestra, s_2.

Y si, sucesivamente, obtenemos una y otra muestra, obtendremos una y otra media muestral, y una y otra desviación típica muestral. Por lo tanto, en realidad, lo que estamos viendo es que la media y la varianza muestrales (y en general, cualquier parámetro de una muestra aleatoria simple) son, en realidad, variables aleatorias que, como tales, deben tener su distribución, su media, su varianza...

Vamos a recordar dos definiciones que ya introdujimos al comienzo del curso.

Un **parámetro muestral** es un parámetro (media, varianza, ...) referido a una muestra de una variable aleatoria.
Un **parámetro poblacional** es un parámetro (media, varianza, ...) referido a la distribución poblacional de una variable aleatoria.

Pues bien, asociados a estos dos conceptos tenemos ahora las siguientes definiciones.

La **distribución en el muestreo** de un parámetro muestral es su distribución de probabilidad.
El **error estandar** de un parámetro muestral es la desviación típica de su distribución en el muestreo.

El problema es que, en general, es bastante difícil conocer la distribución en el muestreo de los parámetros muestrales.

Sin embargo, el caso en el que resulta más sencillo hacerlo es probablemente el más importante. Como vamos a ver, si la variable que observamos sigue una distribución normal, podremos conocer de forma exacta las distribuciones en el muestreo de los dos parámetros más importantes, la media y la varianza.

¿Y si la variable no es normal? Si lo que pretendemos es estudiar la media y la varianza muestrales, recordemos que el Teorema Central del Límite nos dice que si una variable es suma de otras variables, su distribución es aproximadamente normal, y la media es suma de las variables de la muestra. Es decir, si la variable no es normal, todavía podemos tener confianza de que lo que hagamos para variables normales puede ser válido.

6.4. Distribuciones en el muestreo relacionadas con la distribución normal

En este apartado simplemente vamos a presentar una serie de resultados acerca de la distribución en el muestreo, es decir, acerca de las distribuciones de probabilidad, de algunos parámetros muestrales que pueden obtenerse asociados a una variable aleatoria normal.

Algunas de estas distribuciones aparecen por primera vez, así que debemos definirlas previamente. Por otra parte, sus funciones de densidad son bastante poco tratables. Esto no es ningún problema hoy en día, gracias al uso que podemos hacer de los ordenadores para cualquier cálculo. Además, para poder trabajar con ellas cuando no tenemos un ordenador a mano, existen tablas que pueden ser impresas en papel con muchos valores de sus funciones de distribución.

Nota. Una de las primeras distribuciones en el muestreo será la χ^2. Recordemos que una distribución χ^2 con n grados de libertad es una distribución Gamma de parámetros $\frac{n}{2}$ y $\frac{1}{2}$.

Si Z es una variable aleatoria normal estandar y S una χ^2 con n grados de libertad, siendo ambas independientes, entonces

$$t = \frac{Z}{\sqrt{S/n}}$$

sigue una distribución llamada t **de student con** n **grados de libertad**.

Si S_1 y S_2 son variables aleatorias con distribución χ^2 con n_1 y n_2 grados de libertad independientes, entonces

$$F = \frac{S_1/n_1}{S_2/n_2}$$

sigue una distribución que se denomina F **con** n_1 **y** n_2 **grados de libertad**.

Con estas definiciones ya podemos dar las distribuciones en el muestreo de algunos parámetros muestrales importantes asociados a la normal:

- Sea $X_1, ..., X_n$ una muestra aleatoria simple de una variable $N(\mu, \sigma)$. Entonces, el parámetro muestral

$$t = \frac{\bar{X} - \mu}{S_{n-1}/\sqrt{n}}$$

  sigue una t de Student con $n-1$ grados de libertad.

- Sea una muestra $X_1, ..., X_n$ una muestra aleatoria simple de una variable $N(\mu, \sigma)$. Entonces, el parámetro muestral

$$\chi^2 = \frac{(n-1) S_{n-1}^2}{\sigma^2}$$

  sigue una χ^2 con $n-1$ grados de libertad.

- Sean $X_1, ..., X_{n_1}$ e $Y_1, ..., Y_{n_2}$ muestras aleatorias simples de variables independientes con distribuciones $N(\mu_1, \sigma)$ y $N(\mu_2, \sigma)$. Entonces, el parámetro muestral

$$t = \frac{\left(\bar{X} - \bar{Y}\right) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

  donde

$$S_p^2 = \frac{(n_1 - 1)\left(S_{n-1}^1\right)^2 + (n_2 - 1)\left(S_{n-1}^2\right)^2}{n_1 + n_2 - 2},$$

  sigue una t de Student con $n_1 + n_2 - 2$ grados de libertad.

- Sean $X_1, ..., X_{n_1}$ e $Y_1, ..., Y_{n_2}$ muestras aleatorias simples de variables independientes con distribuciones $N(\mu_1, \sigma)$ y $N(\mu_2, \sigma)$. Entonces, el parámetro muestral

$$\chi^2 = \frac{(n_1 + n_2 - 2) S_p^2}{\sigma^2},$$

  sigue una χ^2 $n_1 + n_2 - 2$ grados de libertad.

- Sean $X_1, ..., X_{n_1}$ e $Y_1, ..., Y_{n_2}$ muestras aleatorias simples de variables independientes con distribuciones

$N(\mu_1, \sigma)$ y $N(\mu_2, \sigma)$. Entonces, el parámetro muestral

$$F = \frac{\left(S_{n-1}^1\right)^2 / \sigma_1^2}{\left(S_{n-1}^2\right)^2 / \sigma_2^2}$$

sigue una distribución F con $n_1 - 1$ y $n_2 - 1$ grados de libertad.

Capítulo 7

Estimación de parámetros de una distribución

¡Datos, datos, datos! -gritó impacientemente-. No puedo hacer ladrillos sin arcilla.

Sherlock Holmes (A. C. Doyle), en *Las aventuras de los bombachos de cobre*

Resumen. Se describen las técnicas más usuales para estimar la media, la varianza y otros parámetros poblacionales mediante valores aislados (estimación puntual) o mediante intervalos de confianza.

Palabras clave: estimador puntual, método de los momentos, método de máxima verosimilitud, intervalo de confianza, nivel de confianza.

7.1. Introducción

En Estadística hay tres formas de inferir un valor a un parámetro de una población:

- Estimando el valor concreto de ese parámetro.
- Estimando una región de confianza para el valor del parámetro.
- Tomando una decisión sobre un valor hipotético del parámetro.

Ejemplo. El rendimiento de un equipo de trabajo en una cadena de producción puede estar representado por el número medio de componentes producidas. Supongamos que un ingeniero pretende proporcionar información acerca de este promedio en su equipo. Existen varias posibilidades:

- Podría simplemente tratar de estimar el promedio de componentes producidas a través de un único valor estimado.
- Podría proporcionar un intervalo de valores en el que tenga mucha confianza que se encuentra el valor promedio.

- Podría comparar el valor promedio de su equipo con un valor hipotético para, por ejemplo, demostrar a la empresa que tiene un mejor rendimiento que el promedio general de la empresa.

En este capítulo nos centraremos en la primera y la segunda forma, que consisten en proporcionar un valor que creemos que está cerca del parámetro (estimación puntual) o en proporcionar un intervalo en el que confiamos que se encuentra el parámetro desconocido (estimación por intervalos de confianza). La tercera posibilidad se estudiará en el capítulo de contrastes de hipótesis.

7.2. Estimación puntual

7.2.1. Definición y propiedades deseables de los estimadores puntuales

Un **estimador puntual,** $\hat{\theta}$, es una regla que nos dice cómo calcular una estimación numérica de un parámetro poblacional desconocido, θ, a partir de los datos de una muestra. El número concreto que resulta de un cálculo, para una muestra dada, se denomina **estimación puntual**.

Ejemplo. Si deseamos obtener estimaciones de la media de una variable aleatoria, lo que parece más lógico sería utilizar como estimador la media muestral. Cada media muestral de cada muestra sería una estimación puntual de la media poblacional.

¿Qué sería deseable que le pasara a cualquier estimador? ¿Qué buenas propiedades debería tener un buen estimador? Vamos a ver dos de ellas.

En primer lugar, parece lógico pensar que si bien el estimador no proporcionará siempre el valor exacto del parámetro, al menos deberá establecer estimaciones que *se equivoquen* en igual medida por exceso que por defecto. Este tipo de estimadores se denominan *insesgados*.

Un estimador $\hat{\theta}$ de un parámetro θ se dice **insesgado** si

$$E\left[\hat{\theta}\right] = \theta.$$

Se denomina **sesgo de un estimador** a $\left|E\left[\hat{\Theta}\right] - \theta\right|$.

Observemos que para comprobar si un estimador es insesgado, en principio es necesario conocer su distribución en el muestreo, para poder calcular su esperanza matemática.

Además de la falta de sesgo, nos gustaría que la distribución de muestreo de un estimador tuviera poca varianza, es decir, que la dispersión de las estimaciones con respecto al valor del parámetro poblacional, fuera baja.

En este sentido, se define el **error estandar de un estimador** como la desviación típica de dicho estimador, y se nota *s.e.*

El **estimador insesgado de mínima varianza** de un parámetro θ es el estimador $\hat{\theta}$ que tiene la varianza más pequeña de entre todos los estimadores insesgados.

Hay que decir que no siempre es fácil encontrar este estimador, y que en ocasiones se admite un ligero sesgo con tal que la varianza del estimador sea mínima.

7.2.2. Estimación de la media de una v.a. La media muestral

Sea una v.a. X, y una muestra aleatoria suya, $X_1, ..., X_N$. Entonces, la media muestral,

$$\bar{X} = \frac{X_1 + ... + X_N}{N}$$

es un estimador insesgado de $E[X]$ y su error estandar es

$$s.e.(\bar{X}) = \frac{\sigma_X}{\sqrt{N}}.$$

El resultado establece algo que podía haberse intuido desde la definición de la media o esperanza matemática de una distribución de probabilidad: si tenemos unos datos (*mas*) de una v.a., una estimación adecuada de la media de la v.a. es la media de los datos.

Hay que tener mucho cuidado con no confundir la media de la v.a., es decir, la media poblacional, con la media de los datos de la muestra, es decir, con la media muestral.

Por otra parte, el error estandar hace referencia a σ_X, que es un parámetro poblacional y, por lo tanto, desconocido. Lo que se suele hacer es considerar la desviación típica muestral como una aproximación de la poblacional para evaluar este error estandar.

7.2.3. Estimación de la varianza de una v.a. Varianza muestral

Sea una v.a. X y una muestra aleatoria simple suya, $X_1, ..., X_N$. Entonces, la varianza muestral,

$$S^2_{X,N-1} = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2}{N-1}$$

es un estimador insesgado de $Var[X]$.

Nota. Al hilo del comentario previo que hicimos sobre la media muestral como estimador *natural* de la media, ahora quizá sorprenda que en el denominador de la varianza muestral aparezca $N-1$ y no N. En este sentido, si consideramos el estimador

$$S^2_{X,N} = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2}{N},$$

se trataría de un estimador no insesgado. A este estimador de la varianza se le conoce habitualmente como **cuasivarianza muestral.** Ojo, hay que advertir que en algunos libros la manera de nombrar a la varianza y a la cuasivarianza muestrales es justo al contrario.

Nota. El que la varianza muestral, S_{N-1}^2, sea un estimador insesgado de la varianza, σ^2, no implica que la desviación típica muestral, $S_{N-1} = \sqrt{S_{N-1}^2}$, sea un estimador insesgado de σ, pero en este caso sí ocurre así.

Ejemplo. Mediante R hemos generado una muestra aleatoria simple de 1000 valores de una distribución $N(0,1)$. Sabemos, por tanto, que la media (poblacional) de los datos es 0 y que la varianza (poblacional) es 1. No obstante, vamos a suponer que desconocemos de qué distribución proceden los datos y vamos a tratar de *ajustar* una distribución teórica partiendo de los valores de la muestra:

$$\mathbf{x}_{1\times 1000} = (-0.9459, -0.9557, 0.2711, 0.2603, 1.014, ...)$$

Para empezar, debemos pensar en una distribución adecuada. Para ello puede observarse el histograma de los datos por si éste recuerda la forma de alguna función de densidad conocida. En este caso, el histograma de la muestra aparece en la Figura 7.1, histograma que recuerda claramente la función de densidad de una distribución normal.

La pregunta inmediata una vez que se opta por ajustar mediante una distribución normal es ¿qué normal? Es decir, ¿qué media y qué varianza se proponen para la distribución que queremos ajustar a estos datos? Una respuesta a esta pregunta la proporcionan los estimadores insesgados que hemos encontrado para estos parámetros. Concretamente,

$$\bar{x} = -0.0133$$

y

$$s_{999} = 0.9813,$$

por lo que ajustaríamos los datos de la muestra $\mathbf{x}$ mediante una distribución

$$N(-0.0133, 0.9813).$$

La densidad de esta distribución aparece también en la Figura 7.1, en trazo continuo, y se observa que ajusta muy bien la forma del histograma.

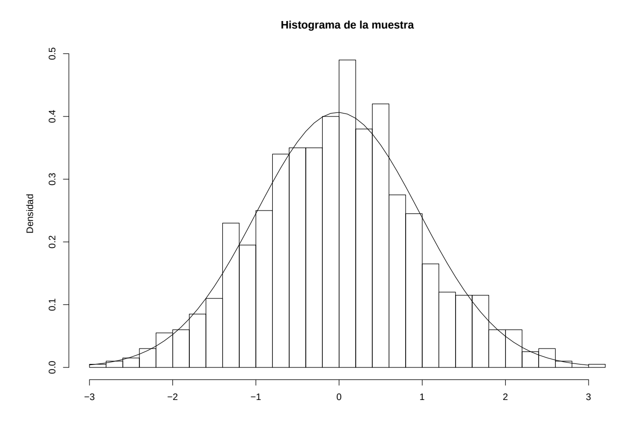


Figura 7.1: Histograma para la muestra $\mathbf{x}_{1\times 1000}$ con 30 intervalos y función de densidad de la distribución $N(-0.0133, 0.9813)$.

7.2.4. Estimación de una proporción poblacional

Supongamos que deseamos estimar una proporción p, desconocida, que representa la probabilidad de un suceso dentro de un espacio muestral. Para ello, se realizan N experimentos asociados al espacio muestral y se cuenta el nº de veces que ocurre ese suceso del cuál queremos estimar su probabilidad, k. En ese caso, la proporción muestral,

$$\hat{p} = \frac{k}{N},$$

es un estimador insesgado de p. Además, su error estandar es

$$s.e.(\hat{p}) = \sqrt{\frac{p(1-p)}{N}}$$

Sobre el error estandar, obsérvese de nuevo que, dado que p es desconocido, en realidad la expresión de $s.e.(\hat{p})$ no puede evaluarse. Sin embargo, es bastante común que si el tamaño de la muestra, N, es grande, se utilice el valor de la estimación, $\hat{p}$, en lugar de p en esa expresión.

De todas formas, obsérvese también que la función $f(p) = p(1-p)$ es menor que $\frac{1}{4}$ si $0 \leq p \leq 1$, luego

$$s.e.(\hat{p}) \leq \sqrt{\frac{1}{4N}} = \frac{1}{2\sqrt{N}}.$$

Es por ello que siempre podemos dar esta cantidad, $\frac{1}{2\sqrt{N}}$, como cota superior del error estandar.

Ejemplo. Si el número de varones en una muestra de 1000 individuos de una población es 507, podemos aproximar la verdadera proporción de varones en toda la población mediante

$$\hat{p} = \frac{507}{1000} = 0.507,$$

con un error estandar por debajo de $\frac{1}{2\sqrt{1000}} = 0.01581139$. La estimación del error estandar de la

estimación sería $\sqrt{0.507 \times 0.493/1000} = 0.01580984$: en este caso, las diferencias son inapreciables.

7.2.5. Obtención de estimadores puntuales. Métodos de estimación

Hasta ahora hemos puesto un ejemplo acerca de la estimación de la media o la varianza de una población mediante la media y la varianza muestral. Sin embargo, nosotros hemos visto muchas distribuciones teóricas que no dependen directamente de la media o la varianza. Por ejemplo, la binomial depende de p, la Gamma de dos parámetros, a y λ, ... ¿Cómo obtener estimadores de estos parámetros?

Existen diversos métodos de estimación de parámetros. Nosotros vamos a ver dos de los más sencillos.

7.2.5.1. Método de los momentos

Vamos a explicar el método sólo para distribuciones de uno o dos parámetros poblacionales, que son las únicas que hemos visto nosotros.

Sea $x_1, ..., x_n$ una muestra de una variable aleatoria X:

1. Si la distribución de X depende de un sólo parámetro, θ, la media poblacional de X, $E[X] = \mu$, será función de θ, $\mu = f(\theta)$. En ese caso, el estimador mediante el método de los momentos de θ, $\hat{\theta}$, se obtiene despejándolo (si es posible) de la ecuación $\bar{x} = f\left(\hat{\theta}\right)$.

2. Si la distribución de X depende de dos parámetros, θ_1 y θ_2, la media poblacional de X, $E[X] = \mu$, será función de ambos, $\mu = f(\theta_1, \theta_2)$ e igualmente la varianza poblacional estará expresada como función de estos parámetros, $VarX = \sigma^2 = g(\theta_1, \theta_2)$. En ese caso, los estimadores mediante el método de los momentos de θ_1 y θ_2, $\hat{\theta}_1$ y $\hat{\theta}_2$, se obtienen despejándolos (si es posible) del sistema de ecuaciones

$$\bar{x} = f\left(\hat{\theta}_1, \hat{\theta}_2\right)$$
$$s_{n-1}^2 = g\left(\hat{\theta}_1, \hat{\theta}_2\right).$$

Ejemplo. En la distribución binomial sabemos que $EX = np$, por lo que $p = \frac{EX}{n}$. Por tanto, dada una muestra de tamaño N de la variable, el método de los momentos propone como estimador de p a

$$\hat{p} = \frac{\bar{x}}{n}.$$

Por cierto, este estimador coincide con el que habíamos considerado en un principio, que era la proporción muestral, es decir, $\hat{p} = k/N$, pero puede haber alguna confusión en la notación. Veamos porqué.

Se supone que tenemos una muestra de tamaño N de datos de una binomial de parámetro n, es decir, tenemos n experimentos, N veces, o sea, un total de $n \times N$ experimentos, con $\sum_i x_i$ éxitos. Luego, en efecto,

$$\hat{p} = \frac{\bar{x}}{n} = \frac{\sum_i x_i}{n \times N},$$

es decir, la proporción muestral, cociente del nº de éxitos entre el nº total de experimentos. No debemos confundirnos con la expresión k/N que pusimos antes porque N no significa lo mismo en ambos casos.

Ejemplo. En la distribución geométrica sabemos que $EX = \frac{1}{p} - 1$, de donde $p = \frac{1}{1+EX}$, luego el método de los momentos propone como estimador a

$$\hat{p} = \frac{1}{1+\bar{x}}.$$

Ejemplo. En el caso de la binomial negativa tenemos dos parámetros. Se sabe que

$$EX = \frac{a(1-p)}{p}$$
$$VarX = \frac{a(1-p)}{p^2}$$

De esta expresión debemos despejar a y p. Dado que

$$\frac{EX}{VarX} = p,$$

se tiene que

$$a = EX \times \frac{p}{1-p} = EX \times \frac{\frac{EX}{VarX}}{1-\frac{EX}{VarX}} = \frac{EX^2}{VarX - EX}$$

de donde se proponen como estimadores

$$\hat{p} = \frac{\bar{x}}{s^2_{X,N-1}}$$
$$\hat{a} = \frac{\bar{x}^2}{s^2_{X,N-1} - \bar{x}}.$$

7.2.5.2. Método de máxima verosimilitud

Este método obedece a un principio muy lógico: dada una muestra, escojamos como estimaciones aquellos valores de los parámetros que hagan *más creibles, más verosímiles*, los datos de la muestra.

Para desarrollar el método debemos tener en cuenta que si tenemos una muestra aleatoria simple de una variable X, $x_1, ..., x_n$, y la función masa o densidad de la variable es $p(x)$, entonces la función masa o densidad de la muestra es

$$p(x_1, ..., x_n) = p(x_1) ... p(x_n).$$

Esta función masa o densidad representa en cierto modo la *credibilidad* de los datos de la muestra.

Dada una variable aleatoria X con función masa o función de densidad $p(x)$, que depende de uno o dos parámetros, y una muestra aleatoria simple de X, $x_1, ..., x_n$, la verosimilitud de la muestra es la función

$$L = p(x_1) ... p(x_n),$$

función que dependerá de los parámetros desconocidos de la variable.

Dada la verosimilitud de una muestra, L,

si L depende de un sólo parámetro, θ, entonces **el estimador máximo-verosímil de** θ se obtiene resolviendo el problema de máximo siguiente:

$$\hat{\theta} = \arg \left[\max_{\theta} L \right].$$

si L depende de dos parámetros, θ_1 y θ_2, entonces **los estimadores máximo-verosímiles de** θ_1 y θ_2 se obtienen resolviendo el problema de máximo siguiente:

$$\left(\hat{\theta}_1, \hat{\theta}_2 \right) = \arg \left[\max_{\theta_1, \theta_2} L \right].$$

Nota. Dado que el máximo de una función coincide con el máximo de su logaritmo, suele ser muy útil maximizar el logaritmo de la función de verosimilitud en vez de la función de verosimilitud.

Ejemplo. Vamos a calcular el estimador máximo verosímil del parámetro p de una distribución $B(n,p)$ basado en una muestra $x_1, ..., x_N$.

En primer lugar, la función de verosimilitud es

$$\begin{aligned} L_{x_1,...,x_N}(p) &= \prod_{i=1}^{N} \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i} \\ &= \left(\prod_{i=1}^{N} \binom{n}{x_i} \right) p^{\sum_{i=1}^{N} x_i} (1-p)^{nN - \sum_{i=1}^{N} x_i}. \end{aligned}$$

Su logaritmo resulta

$$\ln L_{x_1,...,x_N}(p) = \ln \left(\prod_{i=1}^{N} \binom{n}{x_i} \right) + \left(\sum_{i=1}^{N} x_i \right) \times \ln p + \left(nN - \sum_{i=1}^{N} x_i \right) \ln (1-p).$$

Para maximizar esta función derivamos respecto a p e igualamos a cero:

$$\frac{\sum_{i=1}^{N} x_i}{p} - \frac{nN - \sum_{i=1}^{N} x_i}{1-p} = 0,$$

de donde

$$\frac{p}{1-p} = \frac{\sum_{i=1}^{N} x_i}{nN - \sum_{i=1}^{N} x_i} = \frac{\bar{x}}{n - \bar{x}} = \frac{\frac{\bar{x}}{n}}{1 - \frac{\bar{x}}{n}}.$$

Luego el estimador es

$$\hat{p} = \frac{\bar{x}}{n}.$$

Obsérvese que coincide con el estimador que obtuvimos por el método de los momentos.

Ejemplo. Vamos a calcular el estimador máximo verosímil del parámetro λ de una distribución $\exp(\lambda)$ basado en una muestra $x_1, ..., x_N$.

Función de verosimilitud:

$$L_{x_1,\ldots,x_N}(\lambda) = \prod_{i=1}^{N} \lambda e^{-\lambda x_i} = \lambda^N e^{-\lambda \sum_{i=1}^{N} x_i}.$$

Logaritmo de la función de verosimilitud:

$$\ln L_{x_1,\ldots,x_N}(\lambda) = N \ln \lambda - \lambda \sum_{i=1}^{N} x_i.$$

Para maximizar esta función, derivamos respecto a λ e igualamos a cero:

$$\frac{N}{\lambda} - \sum_{i=1}^{N} x_i = 0,$$

de donde

$$\hat{\lambda} = \frac{N}{\sum_{i=1}^{N} x_i} = \frac{1}{\bar{x}}.$$

De nuevo el estimador máximo verosímil coincide con el proporcionado por el método de los momentos.

Ejemplo. En el caso de la distribución normal, tenemos dos parámetros. Veamos cómo proceder en esta situación. Vamos a preocuparnos por los estimadores de la media y de la varianza:

La función de verosimilitud:

$$L_{x_1,\ldots,x_N}\left(\mu, \sigma^2\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}.$$

Su logaritmo:

$$\ln L_{x_1,\ldots,x_N}\left(\mu, \sigma^2\right) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\left(\sigma^2\right) - \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2}.$$

Debemos maximizar esta función como función de μ y σ^2. Para ello, derivamos respecto de ambas variables e igualamos a cero:

$$\frac{d}{d\mu}\ln L_{x_1,\ldots,x_N}\left(\mu, \sigma^2\right) = \frac{\sum_{i=1}^{N}(x_i-\mu)}{\sigma^2} = 0$$

$$\frac{d}{d\sigma^2}\ln L_{x_1,\ldots,x_N}\left(\mu, \sigma^2\right) = -\frac{N}{2\sigma^2} + \frac{1}{2}\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{\left(\sigma^2\right)^2} = 0$$

De la primera ecuación se sigue

$$\sum_{i=1}^{N}(x_i-\mu) = \sum_{i=1}^{N} x_i - N\mu = 0,$$

de donde

$$\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N} = \bar{x}.$$

Modelo	**Estimadores por el método de los momentos**	**Estimadores por el método de máxima verosimilitud**
$B(n,p)$	$\hat{p} = \frac{\bar{x}}{n}$	$\hat{p} = \frac{\bar{x}}{n}$
$P(\lambda)$	$\hat{\lambda} = \bar{x}$	$\hat{\lambda} = \bar{x}$
$Geo(p)$	$\hat{p} = \frac{1}{1+\bar{x}}$	$\hat{p} = \frac{1}{1+\bar{x}}$
$BN(a,p)$	$\hat{a} = \frac{\bar{x}^2}{s^2_{X,N-1}-\bar{x}},\ \hat{p} = \frac{\bar{x}}{s^2_{X,N-1}}$	Sólo por métodos numéricos
$\exp(\lambda)$	$\hat{\lambda} = \frac{1}{\bar{x}}$	$\hat{\lambda} = \frac{1}{\bar{x}}$
$Gamma(a,\lambda)$	$\hat{a} = \frac{\bar{x}^2}{s^2_{n-1}},\ \hat{\lambda} = \frac{\bar{x}}{s^2_{n-1}}$	Sólo por métodos numéricos
$N(\mu,\sigma)$	$\hat{\mu} = \bar{x},\ \hat{\sigma} = s_{n-1}$	$\hat{\mu} = \bar{x},\ \hat{\sigma} = s_n$

Cuadro 7.1: Estimadores por el método de los momentos y de máxima verosimilitud de los parámetros de las distribuciones más usuales.

De la segunda, sustituyendo en ella μ por $\bar{x}$,

$$\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{(\sigma^2)^2} = \frac{N}{\sigma^2},$$

de donde

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N} = s_n^2.$$

Nota. De nuevo hay que llamar la atención sobre el hecho de que hemos buscado un estimador, de máxima verosimilitud, de σ^2, no de σ. Sin embargo, no es muy difícil demostrar que el estimador de máxima verosimilitud de σ en la distribución normal es la cuasidesviación típica muestral, s_n.

7.2.6. Tabla resumen de los estimadores de los parámetros de las distribuciones más comunes

En toda esta sección, supongamos que tenemos una muestra $x_1, ..., x_N$ de una variable aleatoria X. Los estimadores según el método de los momentos y de máxima verosimilitud de los parámetros según las distribuciones que hemos descrito aparecen en el Cuadro 7.1.

7.3. Estimación por intervalos de confianza

Sea $x_1, ..., x_N$ una muestra de una determinada v.a. X cuya distribución depende de un parámetro desconocido θ. Un **intervalo de confianza** para θ con un **nivel de significación** α, $I(x_1, ..., x_N)$, es un intervalo real que depende de la muestra, pero que no depende de θ tal que

$$P[\theta \in I(x_1, ..., x_N)] = 1 - \alpha.$$

Al valor $1-\alpha$ también se le llama **nivel de confianza**.

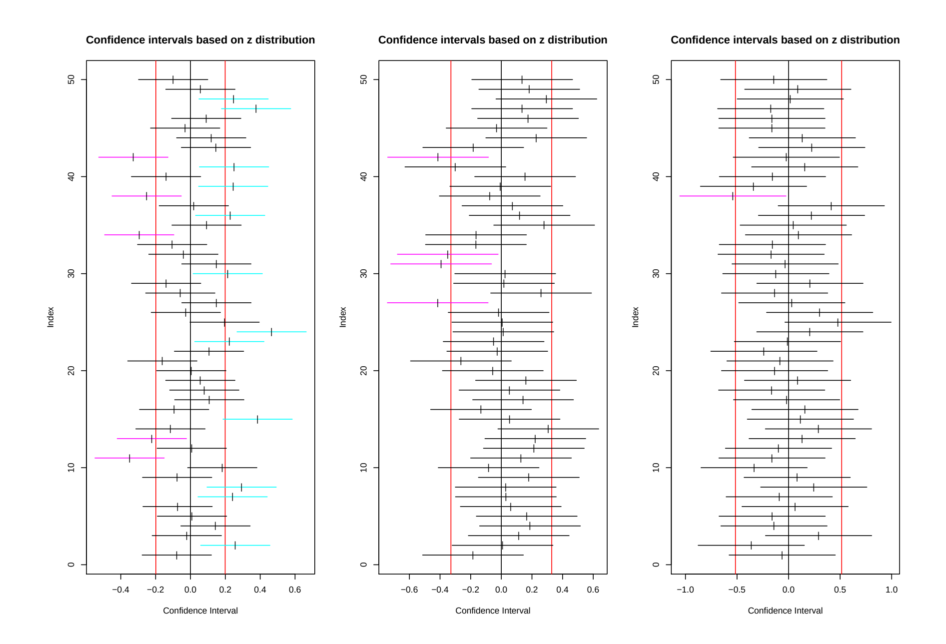


Figura 7.2: Distintos intervalos de confianza para una media a un 68 % (izquierda), a un 90 % (centro) y a un 99 % (derecha). Puede observarse que aumentar el nivel de confianza hace más amplios los intervalos. También puede observarse que no todos los intervalos contienen a la media poblacional (0), pero que el nº de éstos *malos* intervalos disminuye conforme aumentamos el nivel de confianza.

Obsérvese que la filosofía de cualquier intervalo de confianza es proporcionar, basándonos en los datos, una región donde tengamos un determinado nivel de confianza en que el parámetro se encuentra. Como en el caso de los estimadores puntuales, el intervalo de confianza es aleatorio, ya que depende de los datos de una muestra. Además, se da por hecho que existe la posibilidad de que el *verdadero* parámetro θ no quede encerrado dentro del intervalo de confianza, cosa que ocurriría con probabilidad α.

Nota. Al respecto de la interpretación del nivel de confianza, tenemos que decir que, dado que desde el comienzo del curso hemos adoptado una interpretación frecuentista de la probabilidad, un intervalo de confianza al 95 %, por ejemplo, garantiza que si tomamos 100 muestras el parámetro poblacional estará dentro del intervalo en aproximadamente 95 intervalos construidos.

Sin embargo, esta interpretación es absurda en la práctica, porque nosotros no tenemos 100 muestras, sino sólo una.

Nosotros tenemos los datos de una muestra. Con ellos construimos un intervalo de confianza. Y ahora sólo caben dos posibilidades: o el parámetro está dentro del intervalo o no lo está. El parámetro es constante, y el intervalo también. ¡No podemos repetir el experimento! Es por ello que se habla de intervalos *de confianza*, interpretando que tenemos una *confianza* del 95 % en que el parámetro estará dentro.

7.3.1. Intervalos de confianza para la media

Sea X una v.a. con distribución normal de media μ desconocida y varianza σ^2 conocida. Sea una muestra $\mathbf{x} = (x_1, ..., x_N)$ de X, y $\bar{x}$ la media muestral asociada. Entonces,

$$P\left[\mu \in \left[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}}\right]\right] = 1 - \alpha,$$

donde $z_{1-\frac{\alpha}{2}}$[a] es tal que $F_Z\left(z_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$, siendo $Z \to N(0,1)$.

[a]El valor de $z_{1-\frac{\alpha}{2}}$ debe buscarse en la tabla de la normal o calcularse con ayuda del ordenador.

Es decir, la media se encuentra en el intervalo

$$\left[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}}\right]$$

con un $(1-\alpha)\,\%$ de confianza.

No obstante, hay que reconocer que en la práctica es poco probable que se desconozca el valor de la media y sí se conozca el de la varianza, de manera que la aplicación de este teorema es muy limitada. El siguiente resultado responde precisamente a la necesidad de extender el anterior cuando se desconoce el valor de la varianza.

Sea X una v.a. con distribución normal de media μ y varianza σ^2, ambas desconocidas. Sea una muestra $\mathbf{x} = (x_1, ..., x_N)$ de X, la media muestral $\bar{x}$ y la varianza muestral $s^2_{X,N-1}$. Entonces,

$$P\left[\mu \in \left[\bar{x} - t_{1-\frac{\alpha}{2};N-1}\sqrt{\frac{s^2_{X,N-1}}{N}}, \bar{x} + t_{1-\frac{\alpha}{2};N-1}\sqrt{\frac{s^2_{X,N-1}}{N}}\right]\right] = 1 - \alpha,$$

donde $t_{\alpha;N}$[a] es el valor tal que $F_{T_N}(t_{\alpha;N}) = \alpha$, siendo T_N una v.a. con distribución T de Student con N grados de libertad.

[a]El valor de $t_{1-\frac{\alpha}{2}}$ debe buscarse en la tabla de la t o calcularse con ayuda del ordenador

Es decir, confiamos en un $(1-\alpha)\,\%$ en que el intervalo

$$\left[\bar{x} - t_{1-\frac{\alpha}{2};N-1}\sqrt{\frac{s^2_{X,N-1}}{N}}, \bar{x} + t_{1-\frac{\alpha}{2};N-1}\sqrt{\frac{s^2_{X,N-1}}{N}}\right]$$

contiene a la media, que es desconocida.

Ejemplo. Mediante R habíamos simulado 1000 valores de una distribución $N(0,1)$. La media y la desviación típica muestrales de esos 1000 valores resultaron ser $\bar{x} = -0.0133$ y $s_{999} = 0.9813$. Por tanto, el intervalo de confianza que se establece al $95\,\%$ de confianza para la media es

$$\left(-0.0133 \mp 1.96 \frac{0.9813}{\sqrt{1000}}\right) = (-0.074, 0.0475)$$

Obsérvese que, en efecto, la verdadera media, $\mu = 0$, está en el intervalo de confianza.

Los dos resultados que acabamos de enunciar se basan en que se conoce la distribución exacta de la muestra, normal, lo que permite deducir que la media muestral sigue también, y de forma exacta, una distribución normal de media μ y varianza $\frac{\sigma^2}{N}$. Sin embargo, gracias al teorema central del límite se sabe que sea cual sea la distribución de las variables de la muestra aleatoria simple, la media muestral sigue aproximadamente una distribución normal de media μ y varianza $\frac{\sigma^2}{N}$, ya que se obtiene como suma de v.a. independientes con la misma distribución. Por lo tanto, podemos obtener un intervalo de confianza *aproximado* para cualquier media de cualquier distribución, como se recoge en el siguiente resultado.

Sea X una v.a. con distribución cualquiera de media μ, desconocida, y con varianza, σ^2. Sea una muestra $\mathbf{x} = (x_1, ..., x_N)$ de X y la media muestral, $\bar{x}$. Entonces, si N es suficientemente elevado ($N > 30$ es suficiente),

$$P\left[\mu \in \left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{N}}\right]\right] \simeq 1 - \alpha.$$

En esta expresión, si σ es desconocida, puede sustituirse por la desviación típica muestral, s_{n-1}.

Ejemplo. Para dimensionar el tamaño del buffer de un modem ADSL es necesario estimar el promedio de paquetes de datos por milisegundo que recibe el modem.

Se considera que el tiempo (en milisegundos) que transcurre entre paquete y paquete sigue una distribución exponencial de parámetro λ. Obsérvese que la media de esta distribución es $\mu = \frac{1}{\lambda}$, tiempo medio entre paquetes, por lo que λ es precisamente el promedio de paquetes por milisegundo que recibe el modem. Por lo tanto, el objetivo es estimar el parámetro λ, que es el que se utilizará para dimensionar el modem.

Mediante un sniffer acoplado al modem para capturar datos del tráfico, se toman datos de los tiempos entre paquetes de 1001 paquetes, por lo que se tienen 1000 datos de tiempos entre paquetes. La media de estos tiempos resulta ser $\bar{x} = 2.025$, siendo la desviación típica muestral de 1.921.

En primer lugar, vamos a calcular un intervalo de confianza (al 95 %) para la media de la distribución, μ:

$$\left(\bar{x} - z_{0.975}\frac{s_{n-1}}{\sqrt{n}}, \bar{x} + z_{0.975}\frac{s_{n-1}}{\sqrt{n}}\right) = 2.025 \mp 1.96 \times \frac{1.921}{\sqrt{1000}} = (1.906, 2.144).$$

Finalmente, dado que $\lambda = \frac{1}{\mu}$, el intervalo de confianza al 95 % de λ es $\left(\frac{1}{2.144}, \frac{1}{1.906}\right) = (0.466, 0.525)$.

A título informativo, el valor que se considera en el dimensionamiento del modem es un múltiplo (el doble, por ejemplo) del extremo superior del intervalo, en este caso 0.525.

7.3.2. Intervalos de confianza para una proporción

Sea p la probabilidad desconocida de un determinado evento, que llamaremos éxito, que puede ocurrir en un determinado experimento. Supongamos que tenemos una muestra de N realizaciones independientes del experimento, y sea $\hat{p} = \frac{k}{N}$ la proporción de éxitos en la muestra. Entonces, si N es suficientemente elevado ($N > 30$), se tiene que

$$P\left[p \in \left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{N}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{N}}\right)\right] \simeq 1-\alpha.$$

Ejemplo. La Junta de Andalucía pretende implantar un programa de ayuda a familias con familiares dependientes. Dado que la mayor parte de los Servicios Sociales son competencia de los municipios, la Junta proporcionará los medios económicos, pero serán éstos los encargados de ejecutar el programa.

Los Servicios Sociales de cualquier municipio asumen que, por errores inevitables, no todas las familias a las que subvencionan reunen los requisitos exigidos, pero la Junta les responsabiliza de que esto no ocurra en más del 4 % de ellas. Si se supera este porcentaje, penalizará al municipio.

En un municipio se muestrean 200 familias y se detecta que 12 de ellas (6 %) no cumplen las condiciones exigidas. ¿Debe la Junta sancionar al municipio?

Si nos fijamos sólo en el valor de la estimación puntual, 6 %, sí debería hacerlo, pero no sería justo: 12 errores en una muestra de 200 pueden no ser una evidencia suficiente de que el porcentaje superara el 4 %.

Consideremos un un intervalo de confianza para la proporción de errores (5 % de significación) con los datos obtenidos:

$$0.06 \mp 1.96\sqrt{\frac{0.06(1-0.06)}{200}} = (0.027, 0.093).$$

Por tanto, no hay evidencias de que el porcentaje sea superior al 4 % y no debe sancionarse al municipio.

7.3.3. Intervalos de confianza para la varianza

Análogamente, pueden darse intervalos de confianza para la varianza con la media conocida o desconocida, pero sólo cuando la v.a. observada sigue una distribución gaussiana. Ambos casos se recogen en el siguiente resultado.

Sea X una v.a. con distribución gaussiana de media μ (desconocida) y varianza σ^2. Sea una muestra $\mathbf{x} = (x_1, ..., x_N)$ de X y la media muestral $\bar{x}$. Entonces[a] :

$$P\left[\frac{\sum_{i=1}^{N}(X_i - \bar{x})^2}{\chi^2_{1-\frac{\alpha}{2};N-1}} < \sigma^2 < \frac{\sum_{i=1}^{N}(X_i - \bar{x})^2}{\chi^2_{\frac{\alpha}{2};N-1}}\right] = 1-\alpha.$$

[a] El valor de $\chi^2_{\alpha/2;N-1}$ $\chi^2_{1-\alpha/2;N-1}$y debe buscarse en las tablas de la distribución χ^2 u obtenerse mediante el ordenador.

En esta expresión, $\chi^2_{\alpha;N}$ corresponde con aquel valor tal que $F_{\chi^2}\left(\chi^2_{\alpha;N}\right) = \alpha$, donde χ^2 sigue una distribución χ **cuadrado con N grados de libertad.**

Nota. Un intervalo de confianza para la desviación típica puede obtenerse trivialmente como la raiz cuadrada del intervalo de confianza para la varianza.

Ejemplo. En el ejemplo donde consideramos 1000 valores simulados de una $N(0,1)$ teníamos que $\bar{x} = -0.0133$ y $s_{999} = 0.9813$. Por tanto, teniendo en cuenta que

$$\sum_{i=1}^{N} (X_i - \bar{x})^2 = 999 \times s^2_{999},$$

el intervalo de confianza para la varianza al 95 % que proporciona el teorema es

$$\left(\frac{961.9867}{1.0885 \times 10^3}, \frac{961.9867}{913.3010} \right) = (0.8838, 1.0533) .$$

Obsérvese que $\sigma = 1$ pertenece al intervalo de confianza al 95 %.

Puede que alguno de vosotros esté pensando cuál puede ser el interés de las estimaciones puntuales y, sobre todo, mediante intervalos de confianza de la varianza. Probablemente todos tenemos muy claro qué es una media, incluso una proporción, pero quizá se nos escape la importancia práctica del concepto de varianza.

En este sentido, hay que decir que en el ámbito de la Ingeniería la varianza se utiliza muchísimo en lo que se conoce como *control de calidad*. Los japoneses son, en esto, los pioneros y quizá los mejores expertos. A ellos se les atribuye un principio básico del control de calidad en cualquier proceso básico de producción: **la reducción de la varianza es la clave del éxito en la producción**.

Pensemos en cualquier proceso de fabricación genérico. En él se tratará de obtener un producto sujeto a unas especificaciones concretas. Sin embargo, el error inherente a cualquier proceso experimental provocará:

1. Un aumento o una disminución estructurales del producto con respecto a un valor objetivo. Esto podría detectarse como un sesgo en la media de lo producido con respecto al valor objetivo.
2. Unas diferencias más o menos importantes en los productos resultantes, que podrían ser evaluadas mediante la varianza.

De esas dos posibles problemáticas, la más compleja, sin duda es la segunda. Probablemente no es un grave problema *calibrar* la máquina que produce para que la media se sitúe en el valor objetivo, pero será sin duda más complejo modificarla para que produzca de forma más homogénea, reduciendo así la varianza.

7.3.4. Otros intervalos de confianza

Se pueden establecer intervalos de confianza para la diferencia entre las medias de dos variables aleatorias, para la diferencia entre proporciones o para el cociente de varianzas, entre otros parámetros de interés.

Asimismo, se pueden obtener intervalos de confianza *unilaterales* para cualquiera de los parámetros que hemos mencionado, es decir, intervalos acotados sólo a un lado, frente a los intervalos *bilaterales* que hemos visto aquí.

No obstante, no vamos a detallarlos aquí, aunque su interpretación es análoga a la de los intervalos de confianza que hemos visto. Cualquier paquete de software estadístico puede facilitar estos intervalos sin dificultad.

7.4. Resolución del ejemplo de los niveles de plomo

Recordemos que al principio del curso planteábamos un problema que aparece en un artículo publicado en *Journal of Environmental Engineering* en 2002, titulado "Leachate from Land Disposed Residential Construction Waste", en el que se presenta un estudio de la contaminación en basureros que contienen desechos de construcción y desperdicios de demoliciones. Decíamos allí que *De un sitio de prueba se tomaron 42 muestras de lixiado, de las cuales 26 contienen niveles detectables de plomo. Una ingeniera desea obtener a partir de esos datos una estimación de la probabilidad de que una muestra de un basurero contenga niveles detectables de plomo. No obstante, es consciente de que esa estimación estará basada en esa muestra, que es de sólo 42 datos, luego querrá también obtener una estimación del error que está cometiendo al hacer la estimación. Finalmente, se plantea si con la estimación y el error de ésta, podrá obtener un rango donde la verdadera probabilidad se encuentre con un alto nivel de confianza.* Ahora estamos en condiciones de resolver este problema.

En primer lugar, tenemos que obtener una estimación de la proporción de muestras (o probabilidad) que contienen niveles detectables de plomo. Hemos visto que un estimador insesgado de mínima varianza, que además coincide con el estimador de máxima verosimilitud, de la proporción es la proporción muestral. En nuestro caso, por tanto, podemos estimar la proporción en $\hat{p} = \frac{26}{42} = 0.6190$.. Además, podemos estimar el error estándar de esta estimación en $s.e.(\hat{p}) = \sqrt{\frac{0.6190(1-0.6190)}{42}} = 0.0749$ y, en cualquier caso, decir que este error estandar será inferior a $\frac{1}{2\sqrt{42}} = 0.0771$. En resumen, tenemos una estimación del 61.90 % con un error estandar inferior a un 7.71 %.

Por último, en función de esta estimación y de su error estandar, puede afirmar con un 95 % de confianza que el intervalo

$$0.6190 \mp 1.96 \times 0.0749 = (0.4722, 0.7658)$$

contendrá a la verdadera proporción de muestras con niveles detectables de plomo. Esta última afirmación pone de manifiesto que dar un intervalo de confianza con un nivel de significación aceptablemente bajo (5 %) conduce a un intervalo muy amplio, lo que equivale a decir que aún hay bastante incertidumbre con respecto a la proporción que estamos estimando. Por ello, deberíamos recomendarle a la ingeniera que aumente el tamaño de la muestra.

Capítulo 8

Contrastes de hipótesis paramétricas

> La gran tragedia de la ciencia: la destrucción de una bella hipótesis por un antiestético conjunto de datos.
>
> Thomas H. Huxley.

> La Estadística puede probar todo, incluso la verdad.
>
> N. Moynihan

Resumen. En este capítulo explicamos qué se entiende por contraste de hipótesis estadística y aprendemos a realizar contrastes de este tipo a partir de datos, referidos a algún parámetro poblacional desconocido.

Palabras clave: contraste de hipótesis, error tipo I, error tipo II, estadístico de contraste, p-valor, nivel de significación, nivel de confianza.

8.1. Introducción

Como apuntábamos en la introducción del capítulo anterior, las llamadas **pruebas o contrastes de hipótesis** se utilizan para inferir decisiones que se refieren a un parámetro poblacional basándose en muestras de la variable. Vamos a comenzar a explicar el funcionamiento de un contraste de hipótesis con un ejemplo.

Ejemplo. Los científicos recomiendan que para prever el calentamiento global, la concentración de gases de efecto invernadero no debe exceder las 350 partes por millón. Una organización de protección del medio ambiente quiere determinar si el nivel medio, μ, de gases de efecto invernadero en una región cumple con las pautas requeridas, que establecen un límite máximo de 350 partes por millón. Para ello tomará una muestra de mediciones diarias de aire para decidir si se supera el límite, es decir, si $\mu > 350$ o no. Por tanto, la organización desea encontrar apoyo para la hipótesis $\mu > 350$, llamada **hipótesis alternativa**, obteniendo pruebas en la muestra que indiquen que la hipótesis contraria, $\mu = 350$ (o $\mu \leq 350$), llamada **hipótesis nula**, es falsa.

Dicho de otra forma, la organización va a someter a juicio a la hipótesis nula $\mu \leq 350$. Partirá de *su inocencia*, suponiendo que es cierta, es decir, suponiendo que, en principio, no se superan los límites de

presencia de gases de efecto invernadero, y sólo la rechazará en favor de H_1 si hay pruebas evidentes en los datos de la muestra para ello.

La decisión de rechazar o no la hipótesis nula en favor de la alternativa deberá basarse en la información que da la muestra, a través de alguna medida asociada a ella, que se denomina **estadístico de contraste**. Por ejemplo, si se toman 30 lecturas de aire y la media muestral es mucho mayor que 350, lo lógico será rechazar la hipótesis nula en favor de $\mu > 350$, pero si la media muestral es sólo ligeramente mayor que 350 o menor que 350, no habrá pruebas suficientes para rechazar $\mu \leq 350$ en favor de $\mu > 350$.

La cuestión clave es en qué momento se decide rechazar la hipótesis nula en favor de la alternativa. En nuestro ejemplo, en qué momento podemos decir que la media muestral es suficientemente mayor que 350. El conjunto de estos valores del estadístico de contraste, que permiten rechazar $\mu = 350$ en favor de $\mu > 350$ se conoce como **región de rechazo**.

A la luz de este ejemplo, vamos a tratar de definir de forma general los conceptos que acabamos de introducir.

Un **contraste de hipótesis** es una prueba que se basa en los datos de una muestra de una variable aleatoria mediante la cuál podemos rechazar una hipótesis sobre un parámetro de la población, llamada **hipótesis nula** (H_0), en favor de una hipótesis contraria, llamada **hipótesis alternativa** (H_1).

La prueba se basa en una transformación de los datos de la muestra, lo que se denomina **estadístico de contraste**.

Se rechazará la hipótesis nula en favor de la alternativa cuando el valor del estadístico de contraste se sitúe en una determinada región, llamada **región de rechazo**.

La hipótesis H_0 se suele expresar como una igualdad[a], del tipo $H_0 : \theta = \theta_0$, donde θ es un parámetro de una población y θ_0 es un valor hipotético para ese parámetro. Por su parte, H_1 puede tener tener dos formas:
$H_1 : \theta > \theta_0$, en cuyo caso se habla de **contraste unilateral a la derecha** o **de una cola a la derecha** o **de un extremo a la derecha,** o $H_1 : \theta < \theta_0$, en cuyo caso se habla de **contraste unilateral a la izquierda** o **de una cola a la izquierda** o **de un extremo a la izquierda**.
$H_1 : \theta \neq \theta_0$, en cuyo caso se habla de **contraste bilateral** o **de dos colas** o **de dos extremos.**

[a]De todas formas, también es frecuente expresar H_0 como negación exacta de H_1, en cuyo caso sí puede ser una desigualdad no estricta. Matemáticamente no hay diferencias en estas dos posibilidades.

Uno de los aspectos más importantes y que se suele prestar a mayor confusión se refiere a qué hipótesis considerar como H_0 y cuál como H_1. Una regla práctica para hacerlo correctamente puede ser la siguiente:

1. Si estamos intentando probar una hipótesis, ésta debe considerarse como la hipótesis alternativa.

2. Por el contrario, si deseamos desacreditar una hipótesis, debemos incluir ésta como hipótesis nula.

Ejemplo. Para una determinada edificación se exige que los tubos de agua tengan una resistencia media a la ruptura, μ, por encima de 30 kg por centímetro.

- Como primera situación, supongamos que un proveedor quiere facilitar un nuevo tipo de tubo para ser utilizado en esta edificación. Lo que deberá hacer es poner a trabajar a sus ingenieros, que deben realizar una prueba para decidir si esos tubos cumplen con las especificaciones requeridas. En ese caso, deben proponer un contraste que incluya como hipótesis nula $H_0 : \mu \leq 30$ frente a la alternativa $H_1 : \mu > 30$. Si al realizar el contraste de hipótesis se rechaza H_0 en favor de H_1, el tubo podrá ser utilizado, pero si no se puede rechazar H_0 en favor de H_1, no se tienen suficientes garantías sobre la calidad del tubo y no será utilizado.
- Como segunda situación, un proveedor lleva suministrando su tipo de tubo desde hace años, sin que se hayan detectado, en principio, problemas con ellos. Sin embargo, un ingeniero que trabaja para el gobierno controlando la calidad en las edificaciones viene teniendo sospechas de que ese tipo de tubo no cumple con las exigencias requeridas. En ese caso, si quiere probar su hipótesis, el ingeniero deberá considerar un contraste de la hipótesis nula $H_0 : \mu \geq 30$ frente a $H_1 : \mu < 30$. Dicho de otra forma, sólo podrá contrastar su hipótesis si encuentra datos empíricos que permitan rechazar esa hipótesis nula en favor de su alternativa, que demuestren con un alto nivel de fiabilidad que el proveedor que estaba siendo aceptado ahora no cumple con los requisitos.

De hecho, es importantísimo que desde el principio tengamos claro qué tipo de decisiones puede proporcionarnos un contraste de hipótesis. Aunque ya las hemos comentado, vamos a insistir en ellas. Son las dos siguientes:

1. Si el valor del estadístico de contraste para los datos de la muestra cae en la región de rechazo, podremos afirmar **con un determinado nivel de confianza** que los datos de la muestra permiten rechazar la hipótesis nula en favor de la alternativa.
2. Si el valor del estadístico de contraste para los datos de la muestra no cae en la región de rechazo, no podremos afirmar **con el nivel de confianza exigido** que los datos de la muestra permiten rechazar la hipótesis nula en favor de la alternativa.

La clave radica en que entendamos desde el principio que la hipótesis nula carece de confianza. Es asumida sólo como punto de partida, pero será abandonada cuando los datos empíricos muestren evidencias claras en su contra y a favor de la alternativa. La carga de la prueba de hipótesis radica siempre en la hipótesis alternativa, que es la única hipótesis en la que podremos garantizar un determinado nivel de confianza.

8.2. Errores en un contraste de hipótesis

El contraste de una hipótesis estadística implica, por tanto, una toma de decisión, a favor de H_0 o en contra de H_0 y en favor de H_1. Esto implica que podemos equivocarnos al tomar la decisión de dos formas.

Se llama **error tipo I o falso negativo** a rechazar la hipótesis nula cuando es cierta, y su probabilidad se nota por α, llamado **nivel de significación**.

Se llama **nivel de confianza** a la probabilidad de aceptar la hipótesis nula cuando es cierta, es decir, $1 - \alpha$.

		Estado real	
		H_0	H_1
Decisión en el contraste	H_0	Decisión correcta	Error tipo II
	H_1	Error tipo I	Decisión correcta

Cuadro 8.1: Esquematización de los errorres tipo I y tipo II.

Se llama **error tipo II o falso positivo** a aceptar la hipótesis nula cuando es falsa, y su probabilidad se nota por β.

Se llama **potencia** a la probabilidad de rechazar la hipótesis nula cuando es falsa, es decir, $1-\beta$.

¿Cuál de los dos errores es más grave? Probablemente eso depende de cada contraste, pero en general, lo que se pretende es acotar el error tipo I y tratar de minimizar el error tipo II, es decir, tratar de elegir contrastes lo más potentes posibles garantizando que la probabilidad del error tipo I es inferior a un determinado nivel.

Ejemplo. Un fabricante de minicomputadoras cree que puede vender cierto paquete de software a más del 20 % de quienes compran sus computadoras. Se seleccionaron al azar 10 posibles compradores de la computadora y se les preguntó si estaban interesados en el paquete de software. De estas personas, 4 indicaron que pensaban comprar el paquete. ¿Proporciona esta muestra suficientes pruebas de que más del 20 % de los compradores de la computadora adquirirán el paquete de software?

Si p es la verdadera proporción de compradores que adquirirán el paquete de software, dado que deseamos demostrar $p > 0.2$, tenemos que $H_0 : p = 0.2$ y $H_1 : p > 0.2$.

Sea X : número de posibles compradores de la muestra, en cuyo caso, $X \to B(10, p)$. Utilizaremos el valor de X como estadístico del contraste, rechazando H_0 si X es grande.

Supongamos que establecemos como región de rechazo $x \geq 4$. En ese caso, dado que en la muestra $x = 4$, rechazaríamos H_0 en favor de H_1, llegando a la conclusión de que el fabricante tiene razón.

Pero, ¿cuál es el nivel de confianza de este contraste? Calculemos la probabilidad de error tipo I. Para ello, en el Cuadro 8.2 aparece la distribución de probabilidad del estadístico de contraste que hemos elegido, suponiendo que H_0 es cierta, ya que debemos calcular

$$\begin{aligned}
\alpha &= P[\text{Rechazar } H_0|_{H_0 \text{ es cierta}}] = P[X \geq 4|_{p=0.2}] \\
&= 0.08808 + 2.6424 \times 10^{-2} + 5.505 \times 10^{-3} + 7.8643 \times 10^{-4} \\
&+ 7.3728 \times 10^{-5} + 4.096 \times 10^{-6} + 1.024 \times 10^{-7} \\
&= 0.12087,
\end{aligned}$$

luego el nivel de confianza del contraste es del $(1 - 0.12087) \times 100\,\% = 87.913\,\%$. La conclusión sería que **a la luz de los datos podemos afirmar con un 87.913 % de confianza que** $p > 0.2$.

¿Y si queremos un nivel de confianza mayor, es decir, una probabilidad de error tipo I menor? Debemos reducir la región de rechazo. Si ponemos como región de rechazo $x \geq 5$, ya no podremos rechazar H_0 en

x	$\mathbf{P}[X=x]$	
0	$\binom{10}{0}0.2^0 0.8^{10} = 0.10737$	Región de aceptación
1	$\binom{10}{1}0.2^1 0.8^{9} = 0.26844$	
2	$\binom{10}{2}0.2^2 0.8^{8} = 0.30199$	
3	$\binom{10}{3}0.2^3 0.8^{7} = 0.20133$	
4	$\binom{10}{4}0.2^4 0.8^{6} = 0.08808$	Región de rechazo
5	$\binom{10}{5}0.2^5 0.8^{5} = 2.6424 \times 10^{-2}$	
6	$\binom{10}{6}0.2^6 0.8^{4} = 5.505 \times 10^{-3}$	
7	$\binom{10}{7}0.2^7 0.8^{3} = 7.8643 \times 10^{-4}$	
8	$\binom{10}{8}0.2^8 0.8^{2} = 7.3728 \times 10^{-5}$	
9	$\binom{10}{9}0.2^9 0.8^{1} = 4.096 \times 10^{-6}$	
10	$\binom{10}{10}0.2^{10} 0.8^{0} = 1.024 \times 10^{-7}$	

Cuadro 8.2: Función masa del estadístico de contraste suponiendo cierta H_0, es decir, suponiendo que $p = 0.2$.

favor de H_1, ya que $x = 4$. Además, ahora

$$\begin{aligned}\alpha &= 2.6424 \times 10^{-2} + 5.505 \times 10^{-3} + 7.8643 \times 10^{-4} \\ &+ 7.3728 \times 10^{-5} + 4.096 \times 10^{-6} + 1.024 \times 10^{-7} \\ &= 3.2793 \times 10^{-2},\end{aligned}$$

luego el nivel de confianza sería $\left(1 - 3.2793 \times 10^{-2}\right) \times 100\,\% = 96.721\,\%$, y la conclusión sería que **a la luz de los datos no podemos afirmar que $p > 0.2$ con un 96.721 % de confianza**.

El estudio de β es algo más complicado y no lo abordaremos.

8.3. p-valor de un contraste de hipótesis

Históricamente, la forma más común de actuar en un contraste de hipótesis pasa por elegir un nivel de significación (bajo), que determina un límite para el error tipo I que estamos dispuestos a asumir. Ese nivel de significación determina toda la región de rechazo y, examinando si el valor del estadístico cae en ella, podemos concluir si rechazamos o no la hipótesis nula en favor de la alternativa con el nivel de confianza requerido.

Existe, sin embargo, otra forma de actuar que ha tenido un auge enorme desde que las computadoras se han convertido en una herramienta al alcance de cualquiera. Bajo esta forma de actuar, calcularemos el valor del estadístico de contraste y valoraremos cómo es de extremo este valor bajo la distribución en el muestreo de la hipótesis nula. Si es más extremo que el nivel de significación deseado, se rechazará la hipótesis nula en favor de la alternativa. Esta medida de cuán extremo es el valor del estadístico se llama **p-valor**.

8.3.1. Definición de p-valor

De forma general, supongamos que queremos contrastar una hipótesis estadística simple del tipo $H_0 : \theta = \theta_0$, frente a alguna de las alternativas siguientes: $H_1 : \theta \neq \theta_0$, $H_1 : \theta > \theta_0$ o $H_1 : \theta < \theta_0$. Supongamos además

que el contraste se realiza mediante un estadístico que notaremos S, y que el valor del estadístico para la muestra es s.

El **p-valor** asociado al contraste se define como el mínimo nivel de significación con el que la hipótesis nula sería rechazada en favor de la alternativa.

Ejemplo. En el Ejemplo 8.2 hemos visto cómo podemos rechazar la hipótesis nula con un 87.913 % de confianza, pero no con un 96.721 %. Dicho de otra forma, podemos rechazar la hipótesis nula con un nivel de significación del 12.087 %, pero no con un nivel de significación del 3.279 %. Esto implica que el p-valor estará justo entre estos dos últimos valores.

Dado que normalmente se elige como nivel de significación máximo $\alpha = 0.05$, se tiene que la regla de decisión en un contraste con ese nivel de significación, dado el p-valor, sería la siguiente:
Si $p < 0.05$, rechazamos H_0 en favor de H_1 con más de un 95 % de confianza.
Si $p \geq 0.05$, no podemos rechazar H_0 en favor de H_1 con al menos un 95 % de confianza.

Sin embargo, esta regla de decisión, que es la más habitual, es demasiado reduccionista si no se proporciona el valor exacto del p-valor. La razón es que no es lo mismo rechazar una hipótesis con *al menos* un 95 % de confianza si el p-valor es 0.049 que si es 0.001. Hay que proporcionar siempre el p-valor de un contraste, ya que eso permite a cada lector decidir por sí mismo.

En resumen, el p-valor permite utilizar cualquier otro nivel de significación, ya que si consideramos un nivel de significación α:
Si $p < \alpha$, rechazamos H_0 en favor de H_1 con más de un $(1 - \alpha) \times$ % de confianza.
Si $p \geq \alpha$, no podemos rechazar H_0 en favor de H_1 con al menos un $(1 - \alpha) \times$ % de confianza.

Como conclusión, siempre que hagamos un contraste de hipótesis, debemos facilitar el p-valor asociado.

Como nota final sobre el concepto de p-valor, es importante señalar que, al contrario de lo que erróneamente se piensa en demasiadas ocasiones, el p-valor no es la probabilidad de la hipótesis nula. Mucha gente piensa esto porque es cierto que cuando el p-valor es pequeño es cuando se rechaza la hipótesis nula. Sin embargo, para empezar, no tiene sentido plantearnos la *probabilidad* de la hipótesis nula, ya que ésta, o es cierta, o es falsa: desde una perspectiva clásica de la probabilidad, se habla de la probabilidad de un suceso porque a veces ocurre y a veces no, pero en este caso no podemos pensar así, ya que la hipótesis nula o se da o no se da. En realidad, el p-valor lo que da es un indicio de la certidumbre que tenemos, de la confianza en que la hipótesis nula sea verdad, teniendo en cuenta los datos de la muestra. Esta interpretación tiene más que ver con la interpretación subjetiva de la probabilidad de la que hablamos al principio de curso.

Hay que decir que, en relación a esta interpretación subjetiva de la probabilidad, existe una visión de la Estadística, llamada Estadística Bayesiana, en la que el p-valor sí puede entenderse como la probabilidad de la hipótesis nula, pero entendiendo que medimos la probabilidad de la hipótesis nula, no porque pueda ocurrir o no ocurrir en función del azar, sino porque tenemos incertidumbre sobre ella.

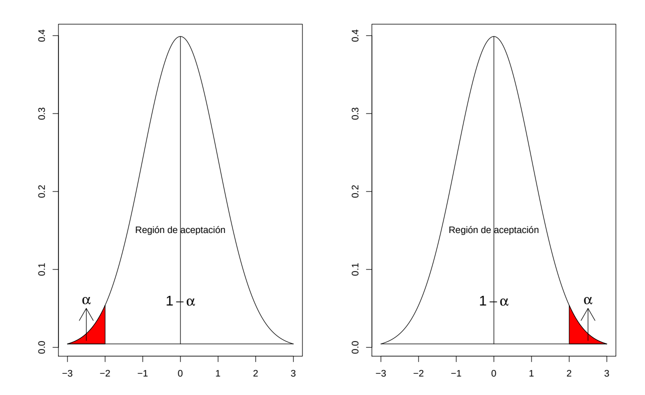


Figura 8.1: Regiones de rechazo en contrastes unilaterales a la izquierda y a la derecha.

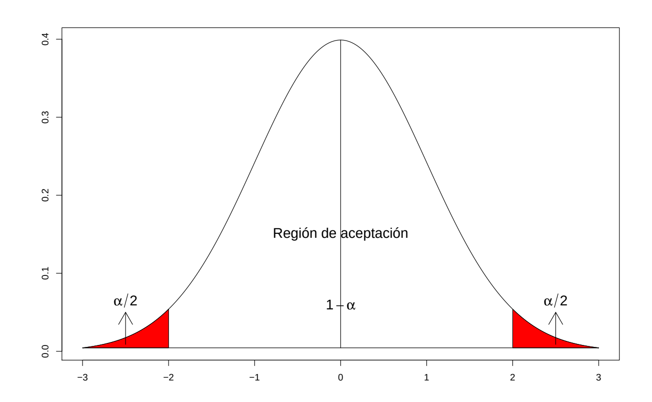


Figura 8.2: Región de rechazo en un contraste bilateral.

8.3.2. Cálculo del p-valor

Para comprender cómo se calcula el p-valor de un contraste es necesario distinguir entre contrastes unilaterales o de una cola frente a contrastes bilaterales o de dos colas.

Como ya comentamos, los contrastes del tipo $H_0 : \theta = \theta_0$, frente a $H_1 : \theta \neq \theta_0$ son **contrastes bilaterales o de dos colas**, ya que el rechazo de la hipótesis nula en favor de la alternativa puede producirse porque el estadístico de contraste toma valores muy altos o muy bajos. Por contra, los contrastes del tipo $H_0 : \theta = \theta_0$, frente a $H_1 : \theta > \theta_0$ o $H_1 : \theta < \theta_0$ son **contrastes unilaterales o de una cola**, ya que el rechazo de la hipótesis nula en favor de la alternativa puede producirse sólo si el estadístico de contraste toma valores muy altos (cuando $H_1 : \theta > \theta_0$, llamado **contraste a la derecha**) o muy bajos (cuando $H_1 : \theta < \theta_0$, **llamado contraste a la izquierda**).

Por tanto, teniendo en cuenta la definición de p-valor, su cálculo se realiza de la siguiente forma:
Si el contraste es unilateral a la izquierda ($H_1 : \theta < \theta_0$),

$$p = P\left[S \leq s/_{H_0}\right].$$

Si el contraste es unilateral a la derecha ($H_1 : \theta > \theta_0$),

$$p = P\left[S > s/_{H_0}\right].$$

Si el contraste es bilateral ($H_1 : \theta \neq \theta_0$),

$$p = 2 \times \text{mín}\left\{P\left[S \leq s/_{H_0}\right], P\left[S > s/_{H_0}\right]\right\}.$$

Hay que decir que el uso del p-valor se ha extendido hasta convertirse en el método más habitual de toma de las decisiones desde que el uso de los ordenadores y de los software de cálculo están a disposición de la mayoría de los usuarios. Hoy en día casi nadie hace Estadística *a mano*, y prácticamente todos los programas estadísticos proporcionan el p-valor como dato para la toma de las decisiones.

En lo que resta del tema lo que vamos a hacer es enunciar distintos contrastes de hipótesis para la media, la varianza o la proporción de una población y para comparar las medias, las varianzas y las proporciones en dos poblaciones distintas. No nos vamos a centrar en los detalles de cómo se deducen sino sólo en cómo se utilizan en la práctica.

De todas formas, es importante hacer una aclaración: cuando los datos proceden de una distribución normal, es muy sencillo obtener la distribución del estadístico del contraste, gracias a los resultados que vimos en el capítulo de distribuciones en el muestreo. Sin embargo, si los datos no proceden de variables normales, esta cuestión es muchísimo más difícil. Afortunadamente, si el tamaño de la muestra es grande, el Teorema Central del Límite garantiza que los parámetros que se basan en sumas basadas en las muestras siguen aproximadamente una distribución normal. Es por ello que en cada tipo de contraste que vamos a describir a continuación se distinguen aquellos que se basan en muestras grandes y los que se basan en muestras reducidas, que sólo podrán ser utilizados si la variable es normal.

En cada caso, vamos a acompañar el contraste con un ejemplo que comentaremos extensamente.

8.4. Contraste para la media de una población

Vamos a suponer que tenemos una muestra $x_1, ..., x_n$ de una variable aleatoria con media poblacional μ. Notaremos $\bar{x}$ a la media muestral y s_{n-1}^2 a la varianza muestral.

8.4.1. Con muestras grandes ($n \geq 30$)

El Cuadro 8.3 incluye un resumen del procedimiento para el contraste. En él, z_p es el valor de una $N(0,1)$ tal que $P[Z < z_p] = p$.

A modo de ejemplo, podemos pensar en que los arqueólogos utilizan el hecho conocido de que los húmeros de los animales de la misma especie tienden a tener aproximadamente las mismas razones longitud/anchura

Tipo de prueba	A la izquierda	Bilateral	A la derecha
Hipótesis	$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$
Estadístico	$z = \frac{\bar{x}-\mu_0}{s_{n-1}/\sqrt{n}}$		
Rechazo	$z < z_\alpha$	$\lvert z\rvert > z_{1-\alpha/2}$	$z > z_{1-\alpha}$
p-valor	$P[Z < z]$	$2P[Z > \lvert z\rvert]$	$P[Z > z]$
Supuestos	$n \geq 30$		

Cuadro 8.3: Contraste para la media con muestras grandes

9.23	10.38	9.76	7.58	9.99	9.46	10.18	9.08	7.09	9.25
12.57	8.71	9.16	10.80	9.86	7.61	8.98	10.81	9.05	9.39
8.42	7.84	9.16	9.40	9.03	9.00	9.25	10.39	8.50	9.51
9.59	8.63	7.48	7.75	8.92	12.85	11.01	8.19	7.44	11.66
11.37	10.06	8.09	9.19	10.79	9.82	9.37	9.66	9.75	9.66

Cuadro 8.4: Datos del ejemplo de las especies

para tratar de discernir si los húmeros fósiles que encuentran en un yacimiento corresponden o no a una nueva especie.

Supongamos que una especie común en la zona donde se enclava un yacimiento, la *Bichus localis*, tiene una razón media longitud/anchura de 9. Los arqueólogos encargados del yacimiento han hallado 50 húmeros fósiles, cuyos datos aparecen en el Cuadro 8.4. ¿Tienen los arqueólogos indicios suficientes para concluir que han descubierto en el yacimiento una especie distinta de la *Bichus localis*?

En primer lugar, observemos que no nos han especificado ningún nivel de significación en el enunciado. En este caso, lo habitual es considerar $\alpha = 0.05$. En caso de que la decisión sea muy relevante, elegiríamos un nivel más bajo.

A continuación debemos plantear las hipótesis del contraste. En principio, la zona de la excavación indica que la especie del yacimiento debería ser la especie *Bichus localis*, salvo que demostremos lo contrario, es decir, la hipótesis nula es $H_0 : \mu = 9$, donde por μ estamos notando la media de la razón longitud/anchura del húmero de la especie del yacimiento. Como hipótesis alternativa nos planteamos que se trate de otra especie, es decir $H_1 : \mu \neq 9$. Se trata, por tanto, de un contraste de dos colas.

Para realizarlo, debemos calcular en primer lugar el estadístico de contraste. Éste, a su vez, requiere del cálculo de la media y de la desviación típica muestral de los datos. Estos valores son, respectivamente, 9.414 y 1.239. Por tanto,

$$z = \frac{9.414 - 9}{1.239/\sqrt{50}} = 2.363.$$

Ahora tenemos que plantearnos si este valor del estadístico nos permite rechazar la hipótesis nula en favor de la alternativa o no. Podemos hacerlo de dos formas:

1. Obteniendo la región de rechazo. Dado que $z_{1-0.05/2} = 1.96$, la región de rechazo es $|z| > 1.96$. Vemos que, en efecto, $2.363 > 1.96$, por lo que podemos rechazar la hipótesis nula en favor de la alternativa con un 95 % de confianza, concluyendo con ese nivel de confianza que se trata de una nueva especie. Nos queda, sin embargo, la duda de saber qué hubiera pasado de tomar un nivel de significación más exigente; por ejemplo, $\alpha = 0.01$.

Tipo de prueba	A la izquierda	Bilateral	A la derecha
Hipótesis	$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$
Estadístico	$t = \frac{\bar{x}-\mu_0}{s_{n-1}/\sqrt{n}}$		
Rechazo	$t < t_{\alpha;n-1}$	$\lvert t\rvert > t_{1-\alpha/2;n-1}$	$t > t_{1-\alpha;n-1}$
p-valor	$P[T_{n-1} < t]$	$2P[T_{n-1} > \lvert t\rvert]$	$P[T_{n-1} > t]$
Supuestos	Distribución de probabilidad aproximadamente normal		

Cuadro 8.5: Contraste para la media con muestras pequeñas

2. Mediante el p-valor. Tenemos que

$$p = 2 \times P[Z > |2.363|] = 0.018.$$

Dado que es inferior al 5 %, podemos rechazar la hipótesis nula en favor de la alternativa con un 95 % de confianza, concluyendo con ese nivel de confianza que la razón media longitud/anchura de los húmeros del yacimiento es distinta de la del *Bichus localis*, pero no podríamos llegar a hacer esa afirmación con un 99 % de confianza (1 % de significación)[1].

8.4.2. Con muestras pequeñas ($n < 30$)

La principal diferencia es que, al no poder utilizar el Teorema Central del Límite por tratarse de muestras pequeñas, debemos añadir como hipótesis la normalidad de los datos. En ese caso, la distribución en el muestreo del estadístico ya no es normal, sino t-student. El resumen aparece en el Cuadro 8.5. En ella, $t_{p;v}$ es el valor de una t de Student con v grados de libertad tal que $P[T_v < t_{p;v}] = p$.

Vamos a aplicar el test en la siguiente situación. El diario Sur publicaba una noticia el 5 de noviembre de 2008 donde se indicaba que *los niveles de concentración de benceno, un tipo de hidrocarburo cancerígeno que se encuentra como vapor a temperatura ambiente y es indisoluble en agua, no superan el máximo permitido por la Directiva Europea de Calidad del Aire, cinco microgramos por metro cúbico. Ésta es la principal conclusión del estudio elaborado por un equipo de la Escuela Andaluza de Salud Pública* en el Campo de Gibraltar. La noticia sólo indicaba que el estudio se basaba en una muestra, dando el valor medio muestral en varias zonas del Campo de Gibraltar, pero no el tamaño ni la desviación típica muestral.

Para realizar el ejemplo, nosotros vamos a imaginar unos datos correspondientes a una muestra de 20 hogares donde se midió la concentración de benceno, arrojando una media muestral de 5.1 microgramos por metro cúbico y una desviación típica muestral de 1.7. Estoy seguro de que, en ese caso, el periódico habría sacado grandes titulares sobre la contaminación por benceno en los hogares del Campo de Gibraltar pero, ¿podemos afirmar que, en efecto, se superan los límites de la Directiva Europea de Calidad del Aire?

En primer lugar, de nuevo no nos indican un nivel de significación con el que realizar la prueba. Escogemos, en principio, $\alpha = 0.05$.

Tenemos que tener cuidado, porque el planteamiento de la prueba, tal y como se nos ha planteado, será contrastar la hipótesis nula $H_0 : \mu = 5$ frente a $H_1 : \mu > 5$, en cuyo caso, un error tipo I se traduce en concluir que se viola la normativa cuando en realidad no lo hace, lo cuál es grave porque genera alarma injustificada en la población, mientras que el error tipo II, el que no controlamos con el α, es concluir que

[1] Debe quedar claro que, estadísticamente, lo que hemos demostrado es que la razón media es distinta de 9. Son los arqueólogos los que deciden que eso implica una nueva especie.

se cumple la normativa cuando en realidad no lo hace, ¡lo cual es gravísimo para la población! Con esto quiero incidir en una cuestión importante respecto a lo que se nos pide que demostremos: se nos dice que nos planteemos si se superan los límites de la normativa, en cuyo caso H_1 debe ser $\mu > 5$, pero en realidad, deberíamos plantearnos la pregunta de si podemos estar seguros de que se está por debajo de los límites máximos permitidos, es decir, deberíamos probar $H_1 : \mu < 5$.

Centrándonos exclusivamente en lo que se nos pide en el enunciado, tenemos que $H_1 : \mu > 5$ determina que se trata de una prueba unilateral a la derecha. El estadístico de contraste es

$$t = \frac{5.1 - 5}{1.7/\sqrt{20}} = 0.263.$$

1. Si queremos concluir con la región de rechazo, ésta está formada por los valores $t > t_{0.95;19} = 1.729$, luego, dado que $0.263 < 1.729$, no podemos afirmar con un 95 % de confianza que se esté incumpliendo la normativa.

2. El p-valor es aún más informativo. Su valor es $p = P[T_{19} > 0.263] = 0.398$, por lo que tendríamos que llegar hasta casi un 40 % de significación para rechazar la hipótesis nula en favor de la alternativa afirmando que se incumple la normativa.

Por lo tanto, tal y como está planteado el problema, no podemos afirmar que se esté incumpliendo la normativa (con un 5 % de significación), por más que un valor muestral de la media, 5.1, parezca indicar que sí. Lo que yo recomendaría a los responsables del cumplimiento la normativa es que aumentaran el tamaño de la muestra, ya que, por ejemplo, si esos mismos datos correspondieran a 1000 hogares en vez de a 20, sí se podría afirmar con un 95 % de confianza que se incumple la normativa.

8.5. Contraste para la diferencia de medias de poblaciones independientes

Sean dos muestras, $x_1, ..., x_{n_1}$ e $y_1, ..., y_{n_2}$, de v.a. independientes con medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2. Sean $\bar{x}$, $\bar{y}$, $\left(s_{n-1}^1\right)^2$ y $\left(s_{n-1}^2\right)^2$ medias y varianzas muestrales.

8.5.1. Con muestras grandes $(n_1, n_2 \geq 30)$

El resumen del procedimiento para el contraste aparece en el Cuadro 8.6.

Vamos a considerar un ejemplo donde aplicar el contraste. Imaginemos que un ingeniero inventa un nuevo método de producción con el que cree que pueden reducirse los tiempos de producción. Para comprobarlo, produce 50 unidades con el nuevo proceso y 30 con el antiguo, contabilizando el tiempo (en segundos) que se tarda en producir cada unidad. En el Cuadro 8.7 aparece un resumen de los resultados.

¿Proporcionan estas muestras pruebas suficientes para concluir que el promedio de tiempo de producción disminuye con el nuevo proceso? Pruébese con $\alpha = 0.05$.

Llamemos μ_1 al tiempo medio de producción bajo el nuevo proceso y μ_2 al tiempo medio de producción bajo el antiguo proceso. Nos piden que contrastemos $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 < \mu_2$ o, lo que es lo mismo, $H_1 : \mu_1 - \mu_2 < 0$: se trata, por tanto, de un test unilateral a la izquierda.

Tipo de prueba	Unilateral a la izquierda	Bilateral	Unilateral a la derecha
Hipótesis	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 < D_0$	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 \neq D_0$	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 > D_0$
Estadístico de contraste	$z = \frac{(\bar{x}-\bar{y})-D_0}{\sqrt{\frac{\left(s^1_{n-1}\right)^2}{n_1}+\frac{\left(s^2_{n-1}\right)^2}{n_2}}}$		
Región de rechazo	$z < z_\alpha$	$\|z\| > z_{1-\alpha/2}$	$z > z_{1-\alpha}$
p-valor	$P[Z < z]$	$2P[Z > \|z\|]$	$P[Z > z]$
Supuestos	$n_1, n_2 \geq 30$. Muestreo independiente y aleatorio		

Cuadro 8.6: Contraste para la diferencia de medias con muestras grandes

Proceso nuevo	Proceso antiguo
$n_1 = 50$	$n_2 = 30$
$\bar{y_1} = 1255$	$\bar{y_2} = 1330$
$s_1 = 215$	$s_2 = 238$

Cuadro 8.7: Datos del ejemplo del nuevo proceso de producción

El estadístico es

$$z = \frac{1255 - 1330}{\sqrt{\frac{215^2}{50} + \frac{238^2}{30}}} = -1.41.$$

Para tomar la decisión podemos obtener la región crítica o el p-valor:

1. La región de rechazo es $z < z_{0.05} = -1.65$. Dado que $z = -1.41$ no cae en esta región, no podemos rechazar la hipótesis nula en favor de la alternativa con $\alpha = 0.05$, es decir, no tenemos un 95 % de confianza en que el nuevo proceso haya disminuido el tiempo medio de producción. No obstante, esta respuesta deja abierta la pregunta, “si no un 95 % de confianza, ¿cuánta?”.

2. Dado que el p-valor es $p = P[Z < -1.41] = 0.079 > 0.05$, no podemos rechazar la hipótesis nula en favor de la alternativa con el nivel de significación $\alpha = 0.05$.

Hay que decir que no hemos podido probar lo que se sospechaba, que el nuevo proceso reducía el tiempo medio de producción, pero los datos apuntan en esta dirección. Desde el punto de vista estadístico, deberíamos recomendar al ingeniero que aumente el tamaño de las muestras porque es posible que en ese caso sí pueda probar esa hipótesis.

8.5.2. Con muestras pequeñas ($n_1 < 30$ *o* $n_2 < 30$) y varianzas iguales

El resumen aparece en el Cuadro 8.8. A propósito de la hipótesis de la igualdad de las varianzas, ésta debe basarse en razones no estadísticas. Lo habitual es que se suponga que son iguales porque el experto que está realizando el contraste tiene razones experimentales para hacerlo, razones ajenas a la estadística.

Vamos a considerar como ejemplo el de un ingeniero que desea comparar dos equipos de trabajo para analizar si se comportan de forma homogénea. Para ello realiza una prueba de destreza entre los trabajadores de ambos equipos: 13 del equipo 1 y 15 del equipo 2, cuyas puntuaciones aparecen en el Cuadro 8.9. ¿Hay indicios suficientes de que existan diferencias entre las puntuaciones medias de los dos equipos? ($\alpha = 0.05$).

Tipo	A la izquierda	Bilateral	A la derecha
Hipótesis	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 < D_0$	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 \neq D_0$	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 > D_0$
Estadístico de contraste	$t = \frac{(\bar{x}-\bar{y})-D_0}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, $s_p^2 = \frac{(n_1-1)\left(s_{n-1}^1\right)^2+(n_2-1)\left(s_{n-1}^2\right)^2}{n_1+n_2-2}$		
Región de Rechazo	$t < t_{\alpha;n_1+n_2-2}$	$\lvert t\rvert > t_{1-\alpha/2;n_1+n_2-2}$	$t > t_{1-\alpha;n_1+n_2-2}$
p-valor	$P[T_{n_1+n_2-2} < t]$	$2P[T_{n_1+n_2-2} > \lvert t\rvert]$	$P[T_{n_1+n_2-2} > t]$
Supuestos	Muestreo independiente y aleatorio. Variables normales. $\sigma_1^2 = \sigma_2^2$		

Cuadro 8.8: Contraste para la igualdad de medias con muestras pequeñas

Equipo 1	59	73	74	61	92	60	84	54	73	47	102	75	33		
Equipo 2	71	63	40	34	38	48	60	75	47	41	44	86	53	68	39

Cuadro 8.9: Datos de las puntuaciones de los dos equipos de trabajo

Nos piden que contrastemos la igualdad de las medias ($H_0 : \mu_1 = \mu_2$), frente a la alternativa $H_1 : \mu_1 \neq \mu_2$, por lo que se trata de un contraste bilateral.

En primer lugar, obtenemos los estadísticos muestrales de ambos equipos. Las medias son, respectivamente, 68.2 y 53.8, mientras que las desviaciones típicas muestrales son 18.6 y 15.8. Con estos valores podemos calcular s_p^2:

$$s_p^2 = \frac{12 \times 18.6 + 14 \times 15.8}{13 + 15 - 2} = 294.09.$$

Con este valor ya podemos calcular el estadístico de contraste:

$$t = \frac{68.2 - 53.8}{\sqrt{294.09(\frac{1}{13} + \frac{1}{15})}} = 2.22.$$

Aunque no hemos dicho nada al respecto, vamos a suponer que las varianzas son iguales. Esto no parece descabellado si admitimos que las condiciones en que trabajan ambos equipos determinan que no debe haber diferencias en la variabilidad de sus puntuaciones. Esta hipótesis debe ser admitida y propuesta por el experto (en este caso, el ingeniero) que maneja los datos.

Para obtener la conclusión, como siempre, vamos a obtener la región de rechazo y valorar el p-valor:

1. La región de rechazo es $|t| > t_{0.975;26} = 2.055$. Dado que $t = 2.22$ cae en esa región, podemos rechazar la igualdad de las medias con un 95 % de confianza.

2. Dado que el p-valor, $p = 2P[T_{26} > 2.22] = 0.035$ es inferior a 0.05, podemos rechazar la igualdad de las medias con un 95 % de confianza. De hecho, podríamos llegar a un 96.5 %.

8.5.3. Con muestras pequeñas, varianzas distintas y mismo tamaño muestral

El resumen del contraste se recoge en el Cuadro 8.10

8.5.4. Con muestras pequeñas, varianzas distintas y distinto tamaño muestral

El resumen aparece en el Cuadro 8.11, donde v se redondea al entero más cercano.

Tipo de prueba	Unilateral a la izquierda	Bilateral	Unilateral a la derecha
Hipótesis	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 < D_0$	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 \neq D_0$	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 > D_0$
Estadístico de contraste	$t = \frac{(\bar{x}-\bar{y})-D_0}{\sqrt{\frac{1}{n}\left(\left(s_{n-1}^1\right)^2+\left(s_{n-1}^2\right)^2\right)}}$		
Región de rechazo	$t < t_{\alpha;2(n-1)}$	$\|t\| > t_{1-\alpha/2;2(n-1)}$	$t > t_{1-\alpha;2(n-1)}$
p-valor	$P[T_{\alpha;2(n-1)} < t]$	$2P[T_{\alpha;2(n-1)} > \|t\|]$	$P[T_{\alpha;2(n-1)} > t]$
Supuestos	Las dos muestras se recogen de forma independiente y aleatoria Ambas variables siguen distribuciones aproximadamente normales Las muestras tienen el mismo tamaño, $n_1 = n_2 = n$		

Cuadro 8.10: Contraste para la igualdad de medias con muestras pequeñas varianzas distintas y mismo tamaño muestral

Tipo de prueba	Unilateral a la izquierda	Bilateral	Unilateral a la derecha
Hipótesis	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 < D_0$	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 \neq D_0$	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 > D_0$
Estadístico de contraste	$t = \frac{(\bar{x}-\bar{y})-D_0}{\sqrt{\frac{\left(s_{n-1}^1\right)^2}{n_1}+\frac{\left(s_{n-1}^2\right)^2}{n_2}}}, v = \frac{\left(\frac{\left(s_{n-1}^1\right)^2}{n_1}+\frac{\left(s_{n-1}^2\right)^2}{n_2}\right)^2}{\frac{\left(\frac{\left(s_{n-1}^1\right)^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{\left(s_{n-1}^2\right)^2}{n2}\right)^2}{n_2-1}}$		
Región de rechazo	$t < t_{\alpha;v}$	$\|t\| > t_{1-\alpha/2;v}$	$t > t_{1-\alpha;v}$
p-valor	$P[T_v < t]$	$2P[T_v > \|t\|]$	$P[T_v > t]$
Supuestos	Las dos muestras se recogen de forma independiente y aleatoria Ambas variables siguen distribuciones aproximadamente normales		

Cuadro 8.11: Contraste para la igualdad de medias con muestras pequeñas, varianzas distintas y distinto tamaño muestral

8.6. Contraste para la diferencia de medias de poblaciones apareadas

Tenemos una misma población en la que seleccionamos una muestra de n individuos. En cada uno de ellos observamos dos variables, X e Y. Estas variables no son independientes: las muestras están **apareadas**, $(x_1, y_1), ..., (x_n, y_n)$. Para comparar ambas variables se considera una nueva variable, $D = X - Y$. Notamos $\bar{d}$ a la media muestral de $x_1 - y_1, ..., x_n - y_n$ y $\left(s_{n-1}^d\right)^2$ a su varianza muestral.

8.6.1. Con muestras grandes ($n \geq 30$)

El resumen aparece en el Cuadro 8.12.

8.6.2. Con muestras pequeñas ($n < 30$)

El resumen aparece en el Cuadro 8.13. Veamos un ejemplo.

Una empresa farmaceútica está investigando un medicamento que reduce la presencia en sangre de un com-

Tipo	A la izquierda	Bilateral	A la derecha
Hipótesis	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 < D_0$	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 \neq D_0$	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 > D_0$
Estadístico	$z = \frac{d - D_0}{s_{n-1}^d / \sqrt{n}}$		
Rechazo	$z < z_\alpha$	$\|z\| > z_{1-\alpha/2}$	$z > z_{1-\alpha}$
p-valor	$P[Z < z]$	$2P[Z > \|z\|]$	$P[Z > z]$
Supuestos	$n \geq 30$		

Cuadro 8.12: Contraste para la igualdad de medias en poblaciones apareadas con muestra grande

Tipo	A la izquierda	Bilateral	A la derecha
Hipótesis	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 < D_0$	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 \neq D_0$	$H_0 : \mu_1 - \mu_2 = D_0$ $H_1 : \mu_1 - \mu_2 > D_0$
Estadístico	$t = \frac{d - D_0}{s_{n-1}^d / \sqrt{n}}$		
Rechazo	$t < t_{\alpha;n-1}$	$\|t\| > t_{1-\alpha/2;n-1}$	$t > t_{1-\alpha;n-1}$
p-valor	$P[T_{n-1} < t]$	$2P[T_{n-1} > \|t\|]$	$P[T_{n-1} > t]$
Supuestos	$D = X - Y$, es aproximadamente normal		

Cuadro 8.13: Contraste para la igualdad de medias en poblaciones apareadas y muestra pequeña

ponente no deseado[2]. Antes de sacarlo al mercado necesita un estudio de casos-controles que demuestre su eficacia.

El estudio de casos controles consiste en encontrar un número determinado de parejas de personas con características fisiológicas parecidas; en este caso, la más importante de estas características sería que las parejas caso-control tengan al inicio del estudio el mismo o muy parecido nivel de presencia en sangre del componente no deseado: en cada una de esas parejas, una actúa como caso, tomando la medicación en estudio, y la otra como control, tomando un producto inocuo llamado placebo. Ninguna de las dos personas, ni siquiera el médico o el farmaceútico que controla el proceso, sabe quién es el caso y quién el control. Sólo quien recopila y analiza los resultados, sin contacto alguno con el paciente, tiene esos datos. Esta metodología se conoce como *doble ciego* y evita que el conocimiento de que se está administrando la medicina provoque un efecto en sí mismo. Los datos aparecen en el Cuadro 8.14.

Un análisis costo-beneficio de la empresa farmaceútica muestra que será beneficioso sacar al mercado el producto si la disminución media del componente perjudicial es de al menos 2 puntos. Realicemos una nueva prueba para ayudar a la compañía a tomar la decisión correcta. Los datos son la disminución de presencia en sangre del componente no deseado después de tomar el medicamento o el placebo.

Empecemos por la notación. Vamos a llamar muestra 1 a la del medicamento y muestra 2 a la del placebo. Con esta notación, nos piden que contrastemos $H_0 : \mu_1 - \mu_2 = 2$ frente a $H_1 : \mu_1 > \mu_2 + 2$, o equivalentemente, $H_1 : \mu_1 - \mu_2 > 2$. En ese caso, el estadístico de contraste es

$$t = \frac{3.21 - 2}{1.134/\sqrt{10}} = 3.375$$

y el p-valor asociado es $p = P[T_9 > 3.375] = 0.004$. Vemos que la significación determina un p-valor inferior, por ejemplo, a $\alpha = 0.05$, por lo que podemos concluir con ese nivel de significación que la mejora es superior, en media, a 2 puntos y, por tanto, el medicamento es rentable.

[2] Podría ser colesterol, ácido úrico, ...

Pareja	Medicamento	Placebo	Diferencia
1	32.10	27.10	5.00
2	36.10	31.50	4.60
3	32.30	30.40	1.90
4	29.50	26.90	2.60
5	34.30	29.90	4.40
6	31.90	28.70	3.20
7	33.40	30.20	3.20
8	34.60	31.80	2.80
9	35.20	33.60	1.60
10	32.70	29.90	2.80

Cuadro 8.14: Datos del ejemplo de la compañía farmaceútica

Tipo de prueba	Unilateral a la izquierda	Bilateral	Unilateral a la derecha
Hipótesis	$H_0 : p = p_0$ $H_1 : p < p_0$	$H_0 : p = p_0$ $H_1 : p \neq p_0$	$H_0 : p = p_0$ $H_1 : p > p_0$
Estadístico de contraste	$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$		
p-valor	$P[Z < z]$	$2P[Z > \lvert z\rvert]$	$P[Z > z]$
Región de rechazo	$z < z_\alpha$	$\lvert z\rvert > z_{1-\alpha/2}$	$z > z_{1-\alpha}$
Supuestos	$np_0, n(1-p_0) \geq 10$		

Cuadro 8.15: Contraste para una proporción

8.7. Contraste para la proporción en una población

En esta ocasión tenemos una población donde una proporción dada presenta una determinada característica, que denominamos *éxito*, y cuya probabilidad es p. Deseamos hacer inferencia sobre esta proporción. Para ello seleccionamos una muestra aleatoria simple de tamaño n y contabilizamos la proporción de éxitos en la muestra, $\hat{p}$. El resumen del contraste aparece en el Cuadro 8.15.

Vamos a considerar un primer ejempo relativo a la relación entre el género y los accidentes de tráfico. Se estima que el 60 % de los conductores son varones. Por otra parte, un estudio realizado sobre los datos de 120 accidentes de tráfico muestra que en ellos el 70 % de los accidentes fueron provocados por un varón conductor. ¿Podemos, con esos datos, confirmar que los hombres son más peligrosos al volante?

Si notamos por p a la proporción de varones causantes de accidentes de tráfico, la pregunta se responderá afirmativamente si logramos contrastar la hipótesis $H_1 : p > 0.6$. El valor del estadístico es

$$z = \frac{0.7 - 0.6}{\sqrt{\frac{0.6\times 0.4}{120}}} = 2.236.$$

Por su parte, la región de rechazo sería $|z| > 1.96$ para un $\alpha = 0.05$, luego en efecto, podemos concluir que la proporción de varones causantes de accidentes es superior a la proporción de varones conductores en general. El p-valor, de hecho, es 0.013.

Vamos a analizar con mucho detalle otro ejemplo sobre igualdad de proporciones. De todas formas, lo que quiero enfatizaros con el ejemplo no está relacionado en sí con el hecho de que se refiera a una proporción.

Una marca de nueces afirma que, como máximo, el 6 % de las nueces están vacías. Se eligieron 300 nueces

al azar y se detectaron 21 vacías. Con un nivel de significación del 5 %, ¿se puede aceptar la afirmación de la marca?

- En primer lugar, pedir un nivel de significación del 5 % es equivalente a pedir un nivel de confianza del 95 % ... ¿sobre qué? Nos preguntan si se puede aceptar la afirmación de la marca **con un nivel de significación del 5 %, es decir, con un nivel de confianza del 95 %**. Eso implica que queremos probar con amplias garantías que la marca no miente, y la única forma de hacerlo es poner su hipótesis ($p < 0.06$) en la hipótesis alternativa. Por tanto, tendríamos $H_0 : p \geq 0.06$ frente a lo que afirma la marca, $H_1 : p < 0.06$.

- Ahora bien, fijémonos que la proporción muestral de nueces vacías es $\hat{p} = 21/300 = 0.07$. Es decir, nos piden que veamos si una proporción muestral de 0.07 da suficiente confianza (95 % para ser exactos) de que $p < 0.06$... ¡No da ninguna! Ni siquiera hace falta hacer el contraste con números. Jamás podremos rechazar la hipótesis nula en favor de la hipótesis de la marca, es decir, en absoluto podemos afirmar lo que dice la marca, $p < 0.06$, con un 95 % de confianza. De todas formas, por si hay algún incrédulo, el estadístico de contraste sería $z = \frac{0.07-0.06}{\sqrt{\frac{0.06\times 0.94}{300}}} = 0.729$. La región de rechazo, dado que es un test a la izquierda, sería $z < z_{0.05} = -1.645$. Como vemos, el valor del estadístico de contraste está en la cola de la derecha y la región de rechazo en la de la izquierda. Por eso decía antes que es imposible rechazar la hipótesis nula en favor de la alternativa, independientemente del nivel de confianza requerido.

- Hasta ahora hemos demostrado que la marca no puede afirmar que la proporción de nueces vacías es inferior al 6 % con un 95 % de confianza. De hecho, no lo puede afirmar con ningún nivel de confianza, porque los datos tomados proporcionan una estimación de 0.07 que va justo en contra de su hipótesis.

- Pero vamos a suponer que nos ponemos "gallitos" y decimos: "*es más, podría demostrar que hay evidencias empíricas que proporcionan un 95 % de confianza en que la compañía miente, siendo en realidad la proporción de nueces vacías superior al 6 %*". Ahora somos nosotros los que afirmamos otra cosa: afirmamos $p > 0.06$ con un 95 % de confianza, lo que equivale a decir que hemos planteado un nuevo contraste de hipótesis en el que $H_0 : p \leq 0.06$ frente a $H_1 : p > 0.06$. Las cuentas están casi hechas, ya que el valor del estadístico de contraste es el mismo, $z = 0.729$, mientras que la región de rechazo es $z > z_{0.95} = 1.645$. Ahora el valor del estadístico, es decir, la información que nos dan los datos (21 de 300 nueces vacías), sí es coherente con la hipótesis alternativa, de ahí que esté en la misma cola que la región de rechazo... ¡pero no cae en ella!. Por lo tanto, no tenemos suficientes evidencias en los datos para rechazar la hipótesis nula en favor de la alternativa con un 95 % de confianza, así que no podemos demostrar con ese nivel de confianza que la marca miente.

- En resumen, aunque parezca paradójico, no tenemos suficientes evidencias en los datos para afirmar que la compañía dice la verdad, pero tampoco para demostrar que miente. La diferencia entre ambas hipótesis radica en que no tenemos ninguna confianza en la afirmación de la compañía, y sí alguna confianza en la afirmación contraria. ¿Cuánta confianza tenemos en la afirmación contraria $p > 0.06$? Ese valor viene dado por el p-valor, $P[Z > 0.729] = 0.233$, que determina que el nivel de confianza en $p > 0.06$ es $(1 - 0.233) \times 100\,\% = 72.9\,\%$.

- Finalmente, alguien podría pensar, "*¿y entonces qué hacemos?*". Desde el punto de vista estadístico lo único que podemos recomendar es aumentar el tamaño de la muestra, es decir, romper más de 300 nueces para tomar la decisión. Aparentemente, la información recogida con 300 nueces parece indicar

Tipo de prueba	Unilateral a la izquierda	Bilateral	Unilateral a la derecha
Hipótesis	$H_0 : p_1 - p_2 = D_0$ $H_1 : p_1 - p_2 < D_0$	$H_0 : p_1 - p_2 = D_0$ $H_1 : p_1 - p_2 \neq D_0$	$H_0 : p_1 - p_2 = D_0$ $H_1 : p_1 - p_2 > D_0$
Estadístico de contraste	$z = \frac{\hat{p}_1 - \hat{p}_2 - D_0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}$		
Región de rechazo	$z < z_\alpha$	$\|z\| > z_{1-\alpha/2}$	$z > z_{1-\alpha}$
p-valor	$P[Z < z]$	$2P[Z > \|z\|]$	$P[Z > z]$
Supuestos	Al menos 10 éxitos y 10 fracasos		

Cuadro 8.16: Contraste para la diferencia de proporciones

que la marca miente. De hecho, si la proporción muestral de 0.07 proviniera de una muestra de 1600 nueces en vez de 300, sí hubiéramos podido demostrar con un 95 % de confianza que la marca miente.

8.8. Contraste para la diferencia de proporciones

En esta ocasión partimos de dos poblaciones dentro de las cuales hay proporciones p_1 y p_2 de individuos con la característica éxito. Pretendemos comparar estas proporciones mediante la toma de muestras de tamaño n_1 y n_2. Notaremos $\hat{p}_1$ y $\hat{p}_2$ las proporciones de éxitos en las muestras. Supondremos de nuevo que las muestras son grandes para poder aplicar el Teorema Central del Límite a la hora de trabajar con el estadístico de contraste. El resumen del contraste aparece en el Cuadro 8.16.

Vamos a considerar un estudio[3] con datos reales, aunque algo anticuados, referente a la relación entre los accidentes de tráfico y el consumo de alcohol, realizado por la DGT en la Comunidad Autónoma de Navarra en 1991.

Se realizaron pruebas de alcoholemia en 274 conductores implicados en accidentes de tráfico con heridos, de los cuales, 88 dieron positivo. Por su parte, la Guardia Civil de Tráfico realizó en la misma zona 1044 controles de alcoholemia al azar, de los cuales 15 dieron positivo.

Lo que la DGT quiere demostrar es que el alcohol es causante de los accidentes de tráfico. Sin embargo, desde el punto de vista estadístico sólo podemos contrastar la hipótesis de que la proporción de positivos en la prueba de alcoholemia es mayor en el grupo de conductores implicados en accidentes de tráfico.

Notemos por p_1 y p_2 a las verdaderas proporciones en el grupo de implicados en accidentes y en el grupo de conductores no implicados. Se nos pide contrastar $H_0 : p_1 = p_2$ frente a $H_1 : p_1 > p_2$. El estadístico de contraste es

$$z = \frac{\frac{88}{274} - \frac{15}{1044}}{\sqrt{\frac{88+15}{274+1044}(1 - \frac{88+15}{274+1044})(\frac{1}{274} + \frac{1}{1044})}} = 904.29.$$

Está claro que el valor del estadístico es bestial, sin necesidad de valorar la región de rechazo, que sería $z > z_{0.95} = 1.645$, luego podemos rechazar la hipótesis nula en favor de la alternativa con, al menos, el 95 % de confianza. El p-valor, $p = P[Z > 904.29] = 0$ indica que la confianza es, de hecho, bastante mayor.

No puedo resistirme a concluir el ejemplo sin recordar que lo que la DGT realmente querrá dar a entender es que el alcohol es el causante de los accidentes de tráfico, pero que eso no puede ser demostrado con el contraste.

[3] http://www.dgt.es/educacionvial/imagenes/educacionvial/recursos/dgt/EduVial/50/40/index.htm

Tipo de prueba	Unilateral a la izquierda	Bilateral	Unilateral a la derecha
Hipótesis	$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 < \sigma_0^2$	$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 \neq \sigma_0^2$	$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 > \sigma_0^2$
Estadístico de contraste	$\chi^2 = \frac{(n-1)s_{n-1}^2}{\sigma_0^2}$		
Rechazo	$\chi^2 < \chi^2_{\alpha;n-1}$	$\chi^2 < \chi^2_{\alpha/2;n-1}$ o $\chi^2 > \chi^2_{1-\alpha/2;n-1}$	$\chi^2 > \chi^2_{1-\alpha;n-1}$
p-valor	$P[\chi^2_{n-1} < \chi^2]$	$2min(P[\chi^2_{n-1} < \chi^2], P[\chi^2_{n-1} > \chi^2])$	$P[\chi^2_{n-1} > \chi^2]$
Supuestos	Distribución de probabilidad aproximadamente normal		

Cuadro 8.17: Contraste para la varianza

8.9. Contraste para la varianza de una población

De nuevo consideremos que tenemos una variable aleatoria X con varianza σ^2 y que tomamos una muestra de tamaño n, cuya varianza muestral notamos por s_{n-1}^2. Vamos a tratar de hacer inferencia sobre σ^2. El problema es que ahora no podemos aplicar el Teorema Central del Límite, por lo que sólo utilizar los contrastes cuando la variable X es normal. $\chi^2_{p;v}$ es el valor de una χ^2 de v grados de libertad tal que $P\left[\chi^2 < \chi^2_{p;v}\right] = p$.

Las empresa Sidel afirma que su máquina de llenado HEMA posee una desviación típica en el llenado de contenedores de 500ml de producto homogéneo inferior a 0.8 gr.[4] Vamos a suponer que el supervisor de control de calidad quiere realizar una comprobación al respecto. Recopila para ello una muestra del llenado de 50 contenedores, obteniendo una varianza muestral de 0.6 ¿Esta información proporciona pruebas suficientes de que la desviación típica de su proceso de llenado es realmente inferior a 0.8gr.?

Planteamos, en primer lugar, las hipótesis del contraste. Se nos pide que contrastemos $H_0 : \sigma = 0.8$ o, equivalentemente, $H_0 : \sigma^2 = 0.64$ frente a la alternativa $H_1 : \sigma^2 < 0.64$. Se trata, por tanto, de un test unilateral a la izquierda. El estadístico de contraste es

$$\chi^2 = \frac{49 \times 0.6}{0.64} = 45.938.$$

Ahora concluimos a través de la región de rechazo (elegimos $\alpha = 0.05$) y del p-valor:

1. Dado que $\chi^2_{0.05;9} = 33.930$, y $\chi^2 = 45.938 > \chi^2_{0.05;9} = 33.930$, no podemos concluir con al menos un 95 % de confianza que, en efecto, la desviación típica de la cantidad de llenado es inferior a 0.8gr.

2. Dado que el p-valor es $p = P[\chi^2_{49} < 45.938] = 0.4$, bastante alto, tenemos muy serias dudas acerca de que, en efecto, la desviación típica sea realmente inferior a 0.8gr.

Ojo: antes de que la empresa Sidel se enfade con nosotros, no olvidemos que los datos son imaginarios: sólo son reales las especificaciones técnicas de $\sigma < 0.8gr$.

8.10. Contraste para el cociente de varianzas

Tenemos dos muestras, $x_1, ..., x_{n_1}$ y $y_1, ..., y_{n_2}$, de dos variables aleatorias independientes con varianzas σ_1^2 y σ_2^2. Notaremos $(s_{n-1}^1)^2$ y $(s_{n-1}^2)^2$ a las varianzas muestrales. De nuevo sólo podremos considerar el contraste

[4] http://www.sidel.com/es/products/equipment/the-art-of-filling/hema-gw

Tipo	Unilateral a la izquierda	Bilateral	Unilateral a la derecha
Hipótesis	$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ $H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1$	$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$	$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ $H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$
Estadístico	$f = \frac{\left(s_{n-1}^1\right)^2}{\left(s_{n-1}^2\right)^2}$		
Rechazo	$f < f_{\alpha;n_1-1,n_2-1}$	$f < f_{\alpha/2;n_1-1,n_2-1}$ o $f > f_{1-\alpha/2;n_1-1,n_2-1}$	$f > f_{1-\alpha;n_1-1,n_2-1}$
p-valor	$P[F_{n_1-1,n_2-1} < f]$	$2min(P[F_{n_1-1,n_2-1} < f], P[F_{n_1-1,n_2-1} > f])$	$P[F_{n_1-1,n_2-1} > f]$
Supuestos	Las dos muestras se recogen de forma independiente y aleatoria Ambas variables siguen distribuciones aproximadamente normales		

Cuadro 8.18: Contraste para el cociente de varianzas

si ambas variables son normales. El resumen del contraste aparece en el Cuadro 8.18. En él, $f_{p;v_1,v_2}$ es el valor de una F de v_1 y v_2 grados de libertad[5] tal que $P[F < f_{p;v_1,v_2}] = p$.

Para practicar sobre el contraste, consideremos que se han realizado 20 mediciones de la dureza en la escala Vickers de acero con alto contenido en cromo y otras 20 mediciones independientes de la dureza de una soldadura producida sobre ese metal. Las desviaciones estándar de las muestras de dureza del metal y de dureza de la soldadura sobre éste fue de $12.06\mu HV$ y $11.41\mu HV$, respectivamente. Podemos suponer que las durezas corresponden a variables normales e independientes. ¿Podemos concluir que la dureza del metal básico es más variable que la dureza medida en la soldadura?

Vamos a llamar a la dureza sobre el acero, X, y a la dureza sobre la soldadura, Y. Se nos pide que contrastemos $H_0 : \sigma_X^2 = \sigma_Y^2$ frente a la alternativa $H_1 : \sigma_X^2 > \sigma_Y^2$ o, equivalentemente, $H_1 : \frac{\sigma_X^2}{\sigma_Y^2} > 1$. Se trata, por tanto, de una prueba unilateral a la derecha. El estadístico de contraste es

$$f = \frac{12.06^2}{11.41^2} = 1.1172.$$

Vamos a tomar un nivel de significación de $\alpha = 0.05$. La región crítica viene delimitada por el valor $f_{0.95;19,19} = 2.168$. Dado que $f = 1.1172 < f_{0.95;19,19} = 2.168$, no podemos concluir al nivel de significación $\alpha = 0.05$ que la dureza del metal básico sea más variable que la dureza medida en la soldadura.

El p-valor, por su parte, es $p = P[F_{19,19} > 1.1172] = 0.4058$.

8.11. Contraste para las medias de más de dos poblaciones independientes. ANOVA

En algunas de las secciones anteriores hemos conseguido contrastes de hipótesis para valorar si existen diferencias significativas entre dos grupos independientes. Lo que nos planteamos aquí es extender estos contrastes para poder comparar no sólo dos sino tres o más grupos. Se da por hecho, por tanto, que existe un **factor** que separa los valores de la variable en varios grupos (dos o más).

Concretamente, supongamos m muestras independientes unas de otras, cada una de ellas con un tamaño n_i[6]. Supongamos también que cada una de las muestras provienen de poblaciones con distribución normal

[5] De cara al uso de las tablas hay una propiedad bastante útil: $f_{p;v_1,v_2} = 1/f_{1-p;v_2,v_1}$

[6] No es necesario, aunque sí deseable, que todas las muestras tengan el mismo tamaño.

de medias μ_i y varianzas todas iguales, σ^2.

Lo que planteamos es contrastar

$$H_0 : \mu_1 = ... = \mu_m$$

frente a

$$H_1 : \text{ no todas las medias son iguales.}$$

Obsérvese que la alternativa no dice *que todas las medias sean distintas* sino tan sólo que al menos dos de ellas sean diferentes.

Denotemos por $x_1^i, ..., x_{n_i}^i$ a la muestra $i-$ésima, y $\bar{x}_i$ y s_{i,n_i-1}^2 a su media y su varianza muestral, con $i = 1, ..., m$.

Este contraste se denomina ANOVA como acrónimo de *Analysis of Variance*, ya que, como vamos a ver, se basa en analizar a qué se debe la variabilidad total que presentan los datos, si al azar o a las diferencias entre las poblaciones de las que proceden las muestras.

Supongamos que *juntamos* todas las muestras, obteniendo una única muestra global de tamaño

$$N = \sum_{i=1}^{m} n_i,$$

y calculamos su media,

$$\bar{x} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} x_j^i}{N}.$$

Ahora, vamos a preguntarnos por las *fuentes de variación de los datos*:

1. En primer lugar, los datos varían globalmente respecto a la media total. Una medida de esta variación es la **suma de los cuadrados totales**,

$$SCT = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(x_j^i - \bar{x}\right)^2.$$

2. Por otro lado, puede haber diferencias entre las medias de cada grupo y la media total. Podemos medir estas diferencias con la **suma de los cuadrados entre-grupos**:

$$SCE = \sum_{i=1}^{m} n_i \left(\bar{x}_i - \bar{x}\right)^2.$$

   Si la hipótesis nula fuera cierta, sólo habría pequeñas diferencias *muestrales* entre las medias de cada muestra, en cuyo caso, la SCE sería pequeña. Si fuera falsa, habría muchas diferencias entre las medias y con respecto a la media total, en cuyo caso SCE sería grande.

3. Por último, debido a la variabilidad inherente a toda muestra, los datos de cada muestra van a variar respecto a su media particular. Como medida de esta variación consideramos la **suma de los cuadrados dentro de los grupos o intra-grupos:**

$$SCD = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(x_j^i - \bar{x}_i\right)^2 = \sum_{i=1}^{m} (n_i - 1) s_{i,n_i-1}^2.$$

La clave en estas consideraciones lo constituye la siguiente igualdad, conocida como **teorema de partición de la varianza**:

$$SCT = SCE + SCD.$$

Teniendo en cuenta este resultado, el ANOVA consiste en ver si SCE es significativamente grande respecto de SCD. Para ello basta considerar que, suponiendo que la hipótesis nula es cierta:

- $\frac{SCT}{\sigma^2}$ sigue una χ^2 con $N-1$ grados de libertad.
- $\frac{SCE}{\sigma^2}$ sigue una χ^2 con $m-1$ grados de libertad.
- $\frac{SCD}{\sigma^2}$ sigue una χ^2 con $N-m$ grados de libertad.

Así, el estadístico de contraste del test es

$$F = \frac{\frac{SCE}{m-1}}{\frac{SCD}{N-m}},$$

que, suponiendo que la hipótesis nula es cierta, sigue una F de Snedecor con $m-1$ y $N-m$ grados de libertad.

Por lo tanto, el test podemos resumirlo de la siguiente forma:

1. Calculamos

   $$\bar{x} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} x_j^i}{N}$$

   y con ella

   $$SCE = \sum_{i=1}^{m} n_i \left(\bar{x}_i - \bar{x}\right)^2 = \sum_{i=1}^{m} n_i \bar{x}_i^2 - N\bar{x}^2.$$

2. Calculamos

   $$SCD = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(x_j^i - \bar{x}_i\right)^2 = \sum_{i=1}^{m} \left(n_i - 1\right) s_{i,n_i-1}^2.$$

3. Calculamos el estadístico del test:

   $$F = \frac{\frac{SCE}{m-1}}{\frac{SCD}{N-m}}.$$

4. Tomamos la decisión:

   a) Si $F \leq F_{m-1,N-m;1-\alpha}$, no rechazamos la hipótesis nula en favor de la alternativa con un nivel de significación α.

   b) Si $F > F_{m-1,N-m;1-\alpha}$, rechazamos la hipótesis nula en favor de la alternativa con un nivel de significación α.

Ejemplo. En un experimento se prepararon flujos de soldadura con 4 composiciones químicas diferentes. Se hicieron 5 soldaduras con cada composición sobre la misma base de acero, midiendo la dureza en la escala de Brinell. El Cuadro 8.19 siguiente resume los resultados.

Vamos a contrastar si existen diferencias significativas entre las durezas, suponiendo que estas siguen distribuciones normales todas ellas con la misma varianza.

Composición	Media muestral	Desviación típica muestral
A	253.8	9.7570
B	263.2	5.4037
C	271.0	8.7178
D	262.0	7.4498

Cuadro 8.19: Datos del ejemplo de ANOVA

En primer lugar, observemos que los tamaños muestrales son iguales: $n_1 = ... = n_4 = 5$.

Por otra parte, tenemos:

$$\bar{x} = \frac{5 \times 253.8 + 5 \times 263.2 + 5 \times 271.0 + 5 \times 262.0}{20} = 262.5$$

$$SCE = 5 \times (253.8 - 262.5)^2 + ... + 5 \times (262.0 - 262.5)^2 = 743.4$$

$$SCD = (5-1)\, 9.7570^2 + ... + (5-1)\, 7.4498^2 = 1023.6.$$

Por tanto,

$$F = \frac{\frac{743.4}{4-1}}{\frac{1023.6}{20-4}} = 3.8734.$$

Por su parte, el valor de $F_{3,16;0.95}$ es 3.2389, de manera que podemos afirmar que existen diferencias significativas entre las durezas de los 4 compuestos, con un 95 % de confianza.

8.12. El problemas de las pruebas múltiples. Método de Bonferroni

¿Qué ocurre si en un estudio tenemos que realizar más de una prueba de hipótesis? Cada prueba lleva consigo un determinado nivel de confianza y, por tanto, una probabilidad de equivocarnos rechazando una hipótesis nula que es cierta (error tipo I). Cuantas más pruebas hagamos, más probabilidades tenemos de cometer un error en la decisión rechazando una hipótesis nula cierta o, dicho de otra forma, menor confianza tendremos.

El método de Bonferroni es uno de los métodos más simples para tratar de corregir este problema asociado a las pruebas múltiples. Se trata de corregir los p-valores de todas las pruebas que se estén realizando simultáneamente, multiplicándolos por el nº total de pruebas, antes de tomar la decisión.

Ejemplo. En Biología Molecular se estudia la relación que puede tener el nivel de expresión de un gen con la posibilidad de padecer un tipo de cáncer. Un investigador consigue analizar el nivel de expresión de 10 genes en una muestra de pacientes y realiza 10 contrastes de hipótesis donde la hipótesis alternativa de cada uno de ellos dice que un gen está relacionado con la posibilidad de padecer ese cáncer. Los p-valores obtenidos son los siguientes:

$$(0.1, 0.01, 0.21, 0.06, 0.32, 0.24, 0.45, 0.7, 0.08, 0.0003)$$

En principio, tendríamos evidencias de que el 2º y el último gen están significativamente relacionados con ese tipo de cáncer. Sin embargo, debemos corregir el efecto de la realización de las 10 pruebas simultáneas.

Aplicando el método de Bonferroni, debemos multiplicar por 10 los p-valores. En ese caso, el segundo gen ya no puede ser considerado estadísticamente significativo para el riesgo de padecer el cáncer ($0.01 \times 10 > 0.05$); por el contrario, dado que $0.0003 \times 10 < 0.05$, el último gen sigue siendo considerado significativamente relacionado con el cáncer.

8.13. Resolución del ejemplo del del diámetro de los cojinetes

Recordemos el planteamiento: *Un ingeniero industrial es responsable de la producción de cojinetes de bolas y tiene dos máquinas distintas para ello. Le interesa que los cojinetes producidos tengan diámetros similares, independientemente de la máquina que los produce, pero tiene sospechas de que está produciendo algún problema de falta de calibración entre ellas. Para analizar esta cuestión, extrae una muestra de 120 cojinetes que se fabricaron en la máquina A, y encuentra que la media del diámetro es de 5.068 mm y que su desviación estándar es de 0.011 mm. Realiza el mismo experimento con la máquina B sobre 65 cojinetes y encuentra que la media y la desviación estándar son, respectivamente, 5.072 mm y 0.007 mm. ¿Puede el ingeniero concluir que los cojinetes producidos por las máquinas tienen diámetros medios significativamente diferentes?*

En este caso, afortunadamente tenemos un tamaño muestral que va a permitir obviar la hipótesis de normalidad. Vemos que se plantea un supuesto que puede ser analizado a través de la media, en concreto, comparando la media de ambas máquinas. Si llamamos X al diámetro de la máquina A e Y al diámetro de la máquina B, tenemos que contrastar $H_0 : \mu_X = \mu_Y$ frente a $H_1 : \mu_X \neq \mu_Y$.

El estadístico de contraste es

$$z = \frac{5.068 - 5.072}{\sqrt{\frac{0.011^2}{120} + \frac{0.007^2}{65}}} = -3.013.$$

El p-valor asociado es $2 \times P[Z < -3.361] = 0.002$, luego tenemos evidencias de que, en efecto, el diámetro medio de ambas máquinas es distinto.

Capítulo 9

Contrastes de hipótesis no paramétricas

> Todos aprendemos de la experiencia, y la lección en esta ocasión es que nunca se debe perder de vista la alternativa.
>
> Sherlock Holmes (A. C. Doyle), en Las Aventuras de Black Peter

Resumen. Continuando con los contraste de hipótesis, presentamos en este capítulo nuevos contrastes que permitirán decidir si un ajuste mediante una distribución teórica es válido y valorar si existe relación entre variables cualitativas.

Palabras clave: bondad de ajuste, test χ^2 de bondad de ajuste, test de bondad de ajuste de Kolmogorov-Smirnoff, test χ^2 de independencia.

9.1. Introducción

Todos los contrastes que hemos descrito en el capítulo anterior se basan, directa o indirectamente (a través del teorema central del límite) en que los datos se ajustan a la distribución normal, haciendo inferencia de una u otra forma sobre sus parámetros. En este capítulo vamos a considerar contrastes que no necesitan de tal hipótesis, por lo que no se enuncian como contrastes sobre algún parámetro desconocido: de ahí que formen parte de los llamados **contrastes no paramétricos** o **contrastes de hipótesis no paramétricas**.

9.2. Contrastes de bondad de ajuste

Gracias a lo estudiado en el apartado correspondiente a la estimación puntual de parámetros ahora somos capaces de ajustar una distribución a unos datos mediante algún método de estimación (momentos, máxima verosimilitud, ...). Sin embargo, hasta ahora no disponemos de ninguna herramienta capaz de *juzgar* si ese ajuste es bueno o malo, o cómo de bueno es. De hecho, en la relación de problemas correspondiente dejamos abierta esta cuestión, ya que sólo pudimos valorar esta *bondad del ajuste* mediante representaciones gráficas, lo que sólo nos dio una visión parcial del problema, que puede ser muy subjetiva.

Los dos contrastes de hipótesis que vamos a describir ahora van a permitir contrastar como hipótesis nula

$$H_0 : \text{la distribución se ajusta adecuadamente a los datos,}$$

Resultado	Observados	Esperados
1	105	100
2	107	100
3	89	100
4	103	100
5	111	100
6	85	100
Total	600	600

Cuadro 9.1: Frecuencias observadas y esperadas en 600 lanzamientos del dado.

frente a la alternativa

$$H_1 : \text{la distribución no se ajusta adecuadamente a los datos,}$$

facilitando además un p-valor que permitirá, además, comparar la bondad de distintos ajustes.

Decir, por último, que aunque estos dos contrastes de hipótesis pueden aplicarse a cualquier tipo de variables están especialmente indicados para variables de tipo discreto o cualitativo en el caso del primero de ellos (test χ^2 de bondad de ajuste) y para variables de tipo continuo en el segundo (test de Kolmogorov-Smirnov).

9.2.1. Test χ^2 de bondad de ajuste

Ejemplo. Supongamos que un tahur del Missisipi quiere probar un dado para ver si es adecuado para jugar honestamente con él. En ese caso, si notamos por p_i a la probabilidad de que en el lanzamiento del dado resulte el valor $i = 1, 2, ..., 6$, el tahur quiere probar la hipótesis

$$H_0 : p_1 = ... = p_6 = \frac{1}{6}$$

frente a la alternativa de H_1 que algún p_i sea distinta de $\frac{1}{6}$.

Para realizar la prueba, lanzará el dado 600 veces, anotando el número de veces que se da cada resultado. Estas cantidades se denominan *frecuencias observadas*.

Por otra parte, si el dado fuera justo (hipótesis H_0), en 600 lanzamientos deberían darse aproximadamente 100 de cada resultado posible. Éstas frecuencias se denominan *frecuencias esperadas*.

El tahur tomará la decisión con respecto al dado a partir de la comparación de las frecuencias observadas y las esperadas (ver Cuadro 9.1). ¿Qué decidirías tú a la luz de esos datos?

A continuación, vamos a describir el test χ^2, que permite realizar pruebas de este tipo. Como hemos comentado en la introducción, con ella podremos *juzgar* ajustes de los que hemos logrado en el capítulo de estimación puntual, pero también podremos utilizarla en ejemplos como el que acabamos de ver, en el que el experto está interesado en contrastar datos experimentales con respecto a una distribución teórica que le resulta de interés.

En primer lugar y de forma más general, supongamos que tenemos una muestra de tamaño N de una v.a. discreta o cualitativa, X, ajustada a un modelo dado por una distribución.

Consideremos una partición del conjunto de valores que puede tomar la variable: $S_1, ..., S_r$. En principio, esta partición podrían ser simplemente todos y cada uno de los valores que toma la variable X, pero, como veremos, es posible que tengamos que agrupar algunos de ellos.

Seguidamente, consideremos la probabilidad, según la distribución dada por el ajuste que queremos evaluar, de cada una de estas partes,

$$p_i = P\left[X \in S_i/_{H_0}\right] > 0.$$

De igual forma, calculemos O_i, el número de observaciones de la muestra que caen en cada conjunto S_i.

La idea del test es comparar el número de observaciones O_i que caen realmente en cada conjunto S_i con el número esperado de observaciones que deberían caer en S_i si el ajuste es el dado por nuestro modelo, que sería $N \times p_i$. Para ello, una medida que compara estas dos cantidades viene dada por

$$D = \sum_{i=1}^{r} \frac{\left(O_i - N \times p_i\right)^2}{N \times p_i}.$$

Si, para una muestra dada, esta v.a. toma un valor d muy alto, indica que los valores observados *no cuadran* con el ajuste que hemos propuesto (con lo cuál se rechazaría la hipótesis nula en favor de la alternativa); si, por el contrario, toma un valor d bajo, indica que nuestro ajuste corresponde bien con los datos de la muestra, por lo que es *aceptable* la hipótesis nula.

El problema final es decidir cuándo el valor de la v.a. D, d, es lo suficientemente alto como para que nos resulte inaceptable el ajuste. Para decidirlo hay que tener en cuenta que cuando N es razonablemente alto y la hipótesis H_0 es cierta, la distribución de probabilidad de D es χ^2 con $r - k - 1$ grados de libertad, es decir,

$$D/_{H_0} \overset{N>>}{\to} \chi^2_{r-k-1},$$

donde k es el número de parámetros que han sido estimados en el ajuste. Teniendo en cuenta este resultado, se calcula bajo esta distribución la probabilidad de que se de un valor todavía más alto que d (el p-valor, por tanto),

$$p = P\left[D > d/_{H_0}\right].$$

Si esta probabilidad es inferior al 5 %, se rechaza la hipótesis nula en favor de la alternativa con un 95 % de confianza. Dicho de otra forma, se acepta la hipótesis nula sólo si el valor de D entra dentro del 95 % de resultados más favorables a ella.

Esquemáticamente, el proceso es el siguiente:

1. Se enuncia el test:

$$H_0: \text{ los datos siguen la distribución dada por nuestro ajuste}$$
$$H_1: \text{ los datos no siguen la distribución dada por nuestro ajuste}$$

2. Si en la muestra se dan los valores $x_1, ..., x_m$, se calculan las frecuencias esperadas según el ajuste propuesto de cada valor x_i, $N \times P[X = x_i]$, $i = 1, ..., m$. Si alguna de estas frecuencias es inferior a 5, se agrupa con alguna de la más cercana hasta que sumen una frecuencia mayor o igual a 5. Se construye así la partición del conjunto de valores posibles para X, $S_1, ...S_r$, cuyas frecuencias esperadas

$\mathbf{x}_i$	0	1	2	3	4	5	6
Frec. obs.	42	28	13	5	7	3	2

Cuadro 9.2: Frecuencias observadas en la muestra de tiempos entre llegadas.

son todas mayores o iguales a 5. En realidad, esto es sólo una recomendación que puede relajarse: si alguna frecuencia esperada es sólo ligeramente inferior a 5, no es especialmente grave.

3. Se calculan las frecuencias observadas de cada S_i, y lo notamos como O_i.

4. Se calcula el estadístico del test en la muestra

$$d = \sum_{i=1}^{r} \frac{(O_i - N \times p_i)^2}{N \times p_i}.$$

5. Se calcula el p-valor asociado al valor del estadístico,

$$p = P\left[D > d/_{H_0}\right],$$

según una distribución χ^2 con $r - k - 1$ grados de libertad.

6. Se toma la decisión (para un nivel de confianza del 95 %):

   a) Si $p < 0.05$, se rechaza la hipótesis nula en favor de la alternativa, con un 95 % de confianza.

   b) Si $p \geq 0.05$, se concluye que no hay evidencias en contra de afirmar que los datos se ajustan a la distribución dada.

Ejemplo. Los datos que se presentan en el Cuadro 9.2 constituyen una muestra aleatoria simple del tiempo en ms. que transcurre entre la llegada de paquetes transmitidos por un determinado protocolo. En la tabla aparecen los valores junto al número de veces que han sido observados en la muestra.

Se sospecha que una distribución geométrica puede ajustar bien esos datos. Vamos a realizar ese ajuste y contrastar si es aceptable mediante el test de la chi-cuadrado.

En primer lugar, para ajustar una distribución geométrica debemos estimar el parámetro de la misma. Vamos a hacerlo de forma sencilla por el método de los momentos. El valor de la media de la distribución es $EX= de donde $p = \frac{1}{1+EX}$. Por tanto, nuestro estimador será

$$\hat{p} = \frac{1}{1+\bar{x}}.$$

Por su parte,

$$\bar{x} = \frac{0 \times 42 + 1 \times 28 + 2 \times 13 + 3 \times 5 + 4 \times 7 + 5 \times 3 + 6 \times 2}{100} = 1.24,$$

luego $

Así pues, deseamos contrastar en qué medida el ajuste de una $Geo\,(0.4464)$ es válido para los datos de la muestra. Es decir, deseamos contrastar $H_0 : X \to Geo\,(0.4464)$ frente a la alternativa $H_1 : X \nrightarrow Geo\,(0.4464)$.

Vamos a calcular cuáles son las probabilidades teóricas según esa distribución de los valores observados en la muestra:

$$P\left[X=0\right] = 0.4464 \times (1-0.4464)^0 = 0.4464$$

$$P\left[X=1\right] = 0.4464 \times (1-0.4464)^1 = 0.2471$$

$$P\left[X=2\right] = 0.4464 \times (1-0.4464)^2 = 0.1368$$

$$P\left[X=3\right] = 0.4464 \times (1-0.4464)^3 = 0.0757$$

$$P\left[X=4\right] = 0.4464 \times (1-0.4464)^4 = 0.0419$$

$$P\left[X=5\right] = 0.4464 \times (1-0.4464)^5 = 0.0232$$

$$P\left[X=6\right] = 0.4464 \times (1-0.4464)^6 = 0.0128$$

$$P\left[X>6\right] = 1-(0.4464+0.2471+0.1368+0.0757+0.0419+0.0232+0.0128) = 0.0159$$

Ahora tenemos que construir la partición de los valores de la variable que, como sabemos, son 0,1,... Hay que tener en cuenta que debemos procurar que las frecuencias esperadas sean superiores o iguales a 5. Como hay 100 observaciones, será necesario agrupar los valores 4 en adelante en un solo conjunto. Vamos a resumir este planteamiento en el Cuadro 9.3 donde, además, aparecen los residuos al cuadrado entre las frecuencias observadas y esperadas, necesarios para calcular el estadístico del test.

El valor de éste se calcula a partir de los resultados de la tabla de la siguiente manera:

$$d = \frac{6.9696}{44.64} + \frac{0.0841}{27.71} + \frac{0.4624}{13.68} + \frac{6.6049}{7.57} + \frac{6.8644}{9.38} = 1.7973.$$

Finalmente, el p-valor se calcula como $P\left[D > 1.7973\right]$, donde D sigue una χ^2_{5-1-1}, es decir, una *Gamma* de parámetros $(5-1-1)/2$ y $1/2$. Por tanto,

$$p-valor = \int_{1.7973}^{\infty} \frac{\frac{1}{2}\left(\frac{1}{2}x\right)^{\frac{3}{2}-1} e^{-\frac{1}{2}x}}{\Gamma\left(\frac{3}{2}\right)} dx = 0.61552.$$

Al ser superior (muy superior, de hecho) a 0.05, podemos afirmar que no hay evidencias en los datos de la muestra en contra de que éstos sigan una distribución $Geo\,(0.4464)$.

x_i	O_i	$N \times p_i$	$(O_i - N \times p_i)^2$
0	42	44.64	$(42 - 44.64)^2 = 6.9696$
1	28	27.71	$(28 - 27.71)^2 = 0.0841$
2	13	13.68	$(13 - 13.68)^2 = 0.4624$
3	5	7.57	$(5 - 7.57)^2 = 6.6049$
≥ 4	12	9.38	$(12 - 9.38)^2 = 6.8644$

Cuadro 9.3: Frecuencias observadas, frecuencias esperadas y residuos.

9.2.2. Test de Kolmogorov-Smirnoff

En este caso el test es aplicable sobre todo a variables de tipo continuo. Se basa en la comparación de la función de distribución teórica propuesta por el modelo cuyo ajuste estamos evaluando con la función de distribución empírica de los datos.

Concretamente, si tenemos $X_1, ..., X_N$ una muestra de una v.a. X, si notamos por $F(x)$ a la función de distribución del modelo propuesto y por $S_N(x)$ a la función de distribución empírica asociada a la muestra, el estadístico que se utiliza para este contraste viene dado por

$$D_N = \underset{x}{Sup} \left| F(x) - S_N(x) \right|.$$

A la hora de calcular este máximo debemos tener en cuenta que la variable x es de tipo continuo.

La hipótesis nula a contrastar es

$$H_0: \text{ los datos de la muestra se ajustan a la distribución dada por } F(x),$$

frente a la hipótesis alternativa

$$H_1: \text{ los datos de la muestra no se ajustan a la distribución dada por } F(x).$$

Se rechazará la hipótesis nula en favor de la alternativa cuando el p-valor asociado al valor que tome D_N sea inferior a 0.05.

Esquemáticamente, el proceso en el desarrollo del test puede resumirse en los siguientes pasos:

1. Ordenamos los valores de la muestra de menor a mayor: $x_{(1)}, ..., x_{(N)}$.
2. Construimos la función de distribución empírica, que en cada valor de la muestra viene dado por $S_N\left(x_{(i)}\right) = \frac{i}{N}$.
3. El valor del estadístico se calcula como

$$d_N = \max_{1 \leq i \leq N} \left\{ \text{máx} \left\{ \left| F\left(x_{(i)}\right) - S_N\left(x_{(i)}\right) \right|, \left| F\left(x_{(i)}\right) - S_N\left(x_{(i-1)}\right) \right| \right\} \right\}.$$

4. Se rechazará la hipótesis nula en favor de la alternativa si $p = P\left[D_N > d_N\right] < 0.05$, con un $(1-p) \times 100\,\%$ de confianza.

La distribución de probabilidad de D_N, necesaria para calcular el p-valor, no es muy conocida. Además, para evaluar esta probabilidad hay que tener en cuenta el número de parámetros de la distribución en el

ajuste. Una metodología adecuada para ello es conocida como Métodos de Monte Carlo, aunque excede los contenidos de estos apuntes. Debo advertir que muchos de los paquetes estadísticos más habituales pueden inducir a error en el cálculo de este p-valor, ya que proporcionan por defecto aquél correspondiente a un ajuste en el que no se estime ningún parámetro en la distribución bajo la hipótesis nula, dando lugar a una sobreestimación de dicho p-valor.

1.4647	0.4995	0.7216	0.1151	0.2717	0.7842	3.9898	0.1967	0.8103	0.4854
0.2333	0.0814	0.3035	1.7358	0.9021	0.0667	0.0868	0.8909	0.1124	0.0512

Cuadro 9.4: Datos de la muestra.

Ejemplo. Los datos que aparecen en el Cuadro 9.4 corresponden al tiempo en sec. entre conexiones a un servidor. Nos planteamos si una distribución exponencial es adecuada para su ajuste.

En primer lugar hemos de decidir cuál es el ajuste propuesto. El estimador máximo verosímil del parámetro λ de una exponencial coincide con el estimador del método de los momentos, $\hat{\lambda} = \frac{1}{m_1}$. En este caso, $\hat{\lambda} = 1/0.6902 = 1.448\,9$.

Para calcular el valor del estadístico del contraste, debemos evaluar la función de distribución de una $\exp(1.4489)$,

$$F(x) = 1 - e^{-1.4489x},\ x \geq 0$$

con la función de distribución empírica. El Cuadro 9.5 muestra ambas funciones de distribución. De ella se deduce que el valor del estadístico de contraste es $0.172\,72$. El p-valor asociado (calculado por Métodos de Monte Carlo con R) toma el valor

$$P[D_{20} > 0.172\,72] = 0.5707.$$

Por tanto, no hay en los datos evidencia en contra de asumir que siguen una distribución $\exp(1.4489)$.

La Figura 9.1 muestra en una vertiente gráfica la bondad del ajuste y el punto donde se alcanza la distancia máxima entre las función de distribución teórica y empírica.

$x_{(i)}$	$F(x_{(i)})$	$\frac{i}{20}$	$\frac{i-1}{20}$	$x_{(i)}$	$F(x_{(i)})$	$\frac{i}{20}$	$\frac{i-1}{20}$
0.0512	7.1499×10^{-2}	0.05	0	0.4854	0.50505	0.55	0.5
0.0667	9.2119×10^{-2}	0.1	0.05	0.4995	0.51506	0.6	0.55
0.0814	0.11125	0.15	0.1	0.7216	0.64849	0.65	0.6
0.0868	0.11818	0.2	0.15	0.7842	0.67897	0.7	0.65
0.1124	0.15029	0.25	0.2	0.8103	0.69089	0.75	0.7
0.1151	0.1536	0.3	0.25	0.8909	0.72496	0.8	0.75
0.1967	0.24798	0.25	0.3	**0.9021**	**0.72938**	**0.85**	0.8
0.2333	0.28682	0.4	0.35	1.4647	0.88023	0.9	0.85
0.2717	0.32542	0.45	0.4	1.7358	0.91914	0.95	0.9
0.3035	0.3558	0.5	0.45	3.9898	0.99691	1	0.95

Cuadro 9.5: Tabla asociada al Test de Kolmogorov-Smirnov.

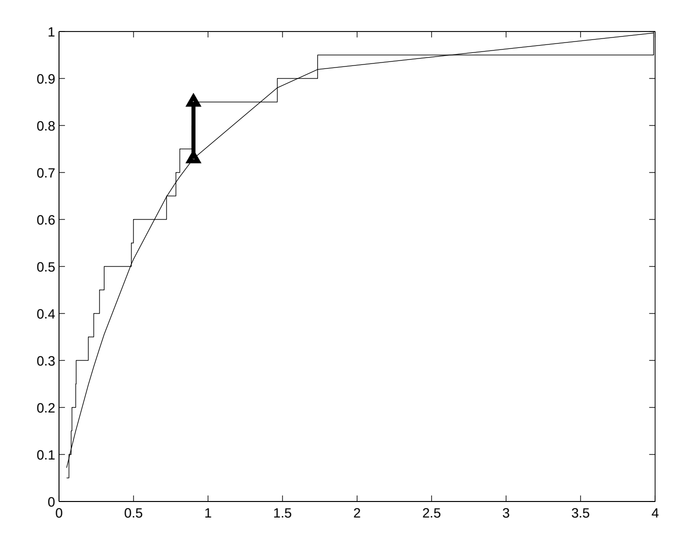


Figura 9.1: Funciones de distribución teórica y empírica. Valor donde se da el estadístico de Kolmogorov-Smirnof.

9.3. Contraste de independencia χ^2

Si nos damos cuenta, desde el capítulo de estadística descriptiva nos hemos centrado exclusivamente en variables de tipo cuantitativo.

Sin embargo, en numerosas ocasiones el objeto de estudio viene determinado, no por una cantidad, sino por una cualidad o un estado no cuantificable. Es por ello que vamos a considerar un contraste relativo a variables de tipo cualitativo, concretamente, para valorar si dos de estas variables están o no significativamente relacionadas.

Ejemplo. ¿Está relacionada la ideología política con el género del votante? Es decir, nos planteamos si el que una persona se declare de izquierdas o de derechas depende de si es varón o mujer. Existen dos variables cualitativas o características que dividen a la población. Lo que nos interesa es si esa división está o no relacionada. ¿Serán más conservadoras las mujeres?

Consideremos en general una población en la que cada individuo se clasifica de acuerdo con dos características, designadas como X e Y. Supongamos que los posibles valores de X son $x_1, ..., x_r$ y los posibles valores de Y son $y_1, ..., y_s$.

Denotemos por p_{ij} a la proporción de individuos de la población cuyas características son simultáneamente x_i e y_j. Denotemos además, como $p_{i.}$ a la proporción de individuos con característica x_i y $p_{.j}$ a la proporción de individuos con característica y_j. En términos de probabilidades, tendremos que si se elige un individuo al azar,

$$P[X = x_i, Y = y_j] = p_{ij}$$

$$P[X = x_i] = p_{i.} = \sum_{j=1}^{s} p_{ij}$$

$$P[Y = y_j] = p_{.j} = \sum_{i=1}^{r} p_{ij}.$$

Lo que pretendemos contrastar es si las dos características son independientes, es decir, si para todo i y para todo j,

$$P[X = x_i, Y = y_j] = P[X = x_i] \times P[Y = y_j],$$

es decir, si

$$p_{ij} = p_{i.} \times p_{.j}.$$

Así pues, podemos enunciar el contraste como

$$H_0 : p_{ij} = p_{i.} \times p_{.j} \text{ para todo } i = 1, ..., r; j = 1, ..., s$$

frente a

$$H_1 : p_{ij} \neq p_{i.} \times p_{.j} \text{ para algún valor de } i \text{ y } j.$$

Para llevar a cabo el contraste tomaremos una muestra de la población de tamaño n. Denotemos por n_{ij} los individuos de esa muestra que toman simultáneamente el valor x_i y el valor y_j (**frecuencias observadas**), $n_{i.} = \sum_{j=1}^{s} n_{ij}$ los individuos de la muestra que toman el valor x_i y $n_{.j} = \sum_{i=1}^{r} n_{ij}$ los que toman el valor y_j.

De esta forma,

$$\hat{p}_{ij} = \frac{n_{ij}}{n}$$

será un estimador basado en la muestra de p_{ij},

$$\hat{p}_{i.} = \frac{n_{i.}}{n}$$

será un estimador basado en la muestra de $p_{i.}$ y

$$\hat{p}_{.j} = \frac{n_{.j}}{n}$$

será un estimador basado en la muestra de $p_{.j}$.

Por otra parte, si la hipótesis nula fuera cierta, el número de individuos en la muestra, de tamaño n, que toman simultáneamente los valores x_i y y_j sería

$$e_{ij} = n \times p_{i.} \times p_{.j}.$$

Basado en la muestra, los valores

$$\begin{aligned} \hat{e}_{ij} &= n \times \hat{p}_{i.} \times \hat{p}_{.j} \\ &= \frac{n_{i.} \times n_{.j}}{n} \end{aligned}$$

(**frecuencias esperadas**) serían sus estimadores.

Finalmente, el estadístico del contraste se basa en comparar los valores reales en la muestra de n_{ij} con los valores $\hat{e}_{ij}$ que se darían si la hipótesis nula fuera cierta, es decir, si las características X e Y fueran

independientes. El valor del estadístico es

$$d = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}.$$

Suponiendo que la hipótesis nula es cierta, la distribución del estadístico del contraste es χ^2 con $(r-1)(s-1)$ grados de libertad, por lo que decidiremos en función del p-valor asociado,

$$p = P\left[D > d/_{H_0}\right],$$

donde $D \to \chi^2_{(r-1)(s-1)}$ o bien:

- Rechazaremos H_0 con nivel de significación α si $d > \chi^2_{(r-1)(s-1);1-\alpha}$.

- No rechazaremos H_0 con nivel de significación α si $d < \chi^2_{(r-1)(s-1);1-\alpha}$.

Hay que hacer una última observación: para que en efecto $D \to \chi^2$ con $(r-1)(s-1)$ es necesario que todas (o casi todas) las frecuencias esperadas $\hat{e}_{ij}$ sean mayores o iguales a 5. Si alguna o algunas de ellas no lo son, la distribución χ^2 podría no ser adecuada y el resultado del test incorrecto. Para que esto no ocurra es recomendable que el tamaño de la muestra sea grande.

Ejemplo. Se toma una muestra de 300 personas, preguntándoles si se consideran más de derechas, más de izquierdas o de centro y anotando su género. El resultado se resume en la siguiente tabla:

	Izquierda	Derecha	Centro	Total
Mujeres	68	56	32	156
Hombres	52	72	20	144
Total	120	128	52	300

Este tipo de tablas se conocen como **tablas de contingencia**. Contiene los valores que hemos notado n_{ij} y, en los márgenes inferior y lateral derecho, los valores $n_{i.}$ y $n_{.j}$.

Vamos a ver si el género está relacionado con la ideología. Si no fuera así, si la ideología fuera independiente del género, se tendría en una muestra de 300 individuos las frecuencias esperadas serían

	Izquierda	Derecha	Centro	Total
Mujeres	$300\frac{156}{300}\frac{120}{300}$	$300\frac{156}{300}\frac{128}{300}$	$300\frac{156}{300}\frac{52}{300}$	156
Hombres	$300\frac{144}{300}\frac{120}{300}$	$300\frac{144}{300}\frac{128}{300}$	$300\frac{144}{300}\frac{52}{300}$	144
Total	120	128	52	300

	Izquierda	Derecha	Centro	Total
Mujeres	62.40	66.56	27.04	156
Hombres	57.60	61.44	24.96	144
Total	120	128	52	300

.

El valor del estadístico de contraste es, por tanto,

$$D = \frac{(68-62.40)^2}{62.40} + \frac{(56-66.56)^2}{66.56} + \frac{(32-27.04)^2}{27.04} + \\ + \frac{(52-57.60)^2}{57.60} + \frac{(72-61.44)^2}{61.44} + \frac{(20-24.96)^2}{24.96} = 6.433.$$

Por su parte, $\chi^2_{(2-1)(3-1);0.95} = 5.991$, de manera que podemos rechazar la hipótesis nula en favor de la alternativa, afirmando con un 95 % de confianza que el genero está relacionado con la ideología. ¿En qué sentido lo estará?

- Si nos centramos sólo en los de izquierdas, tenemos que el porcentaje de hombres y mujeres es de $\frac{68}{120} \times 100\,\% = 56.667\,\%$ y de $\frac{52}{120} \times 100\,\% = 43.333\,\%$, respectivamente.
- Si nos centramos sólo en los de derechas, tenemos que el porcentaje de hombres y mujeres es de $\frac{56}{128} \times 100\,\% = 43.75\,\%$ y de $\frac{72}{128} \times 100\,\% = 56.25\,\%$, respectivamente.
- Finalmente, si nos centramos sólo en los de centro, tenemos que el porcentaje de hombres y mujeres es de $\frac{32}{52} \times 100 = 61.538\,\%$ y de $\frac{20}{52} \times 100 = 38.462\,\%$, respectivamente.

Lo que parece que ocurre es que las mujeres tienen mayor preferencia por la derecha. Sin embargo, esta afirmación no se ha contrastado, sino que se basa simplemente en datos descriptivos[1].

9.4. Resolución del ejemplo de los accidentes laborales

Redordemos el planteamiento: *En una empresa se sospecha que hay franjas horarias donde los accidentes laborales son más frecuentes. Para estudiar este fenómeno, contabilizan los accidentes laborales que sufren los trabajadores según franjas horarias, durante un año. Los resultados aparecen en la tabla.*

Horas del día	Número de accidentes
8-10 h.	47
10-12 h.	52
13-15 h.	57
15-17 h.	63

Con esa información, los responsables de seguridad de la empresa deben decidir si hay franjas horarias donde los accidentes son más probables o si, por el contrario, éstos ocurren absolutamente al azar.

En primer lugar debemos plantearnos la hipótesis que queremos contrastar. El hecho de que ocurran los accidentes absolutamente al azar vendría a decir que la probabilidad de ocurrencia es la misma en cada franja horaria (puesto que todas ellas tienen la misma amplitud). Por ello, si notamos p_i a la probabilidad de que ocurra un accidente en la i-ésima franja horaria, nos planteamos como hipótesis nula $H_0 : p_1 = ... = p_4 = \frac{1}{4}$ frente a la alternativa de que no todas las probabilidades sean iguales.

Para realizar el contraste podemos considerar un contraste de bondad de ajuste en el que la distribución de probabilidad sea una uniforme discreta, que no tiene parámetros.

En este caso, el estadístico de contraste es muy sencillo:

$$\chi^2 = \frac{(47 - 219 \times (1/4))^2}{219 \times (1/4)} + \frac{(52 - 219 \times (1/4))^2}{219 \times (1/4)} + \frac{(57 - 219 \times (1/4))^2}{219 \times (1/4)} + \frac{(63 - 219 \times (1/4))^2}{219 \times (1/4)} = 2.571.$$

Por su parte, el p-valor es $p = P[\chi^2_{4-0-1} > 2.571] = 0.462$, por lo que no tenemos evidencias en estos datos que hagan pensar en que hay franjas horarias más propicias a los accidentes.

Capítulo 10

Regresión lineal simple

> Un político debe ser capaz de predecir lo que pasará mañana, y la semana, el mes y el año próximos. Y también debe ser capaz de explicar por qué no acertó.
>
> Winston Churchill

Resumen. En este capítulo se describe el modelo de regresión lineal simple, que asume que entre dos variables dadas existe una relación de tipo lineal contaminada por un error aleatorio. Aprenderemos a estimar dicho modelo y, a partir de estas estimaciones y bajo determinadas hipótesis, podremos extraer predicciones del modelo e inferir la fortaleza de dicha relación lineal.

Palabras clave: regresión lineal simple, variable dependiente, variable independiente, error aleatorio, nube de puntos, principio de mínimos cuadrados, coeficiente de correlación lineal, coeficiente de determinación lineal, bondad del ajuste, predicción, estimación.

10.1. Introducción

Uno de los aspectos más relevantes que aborda la Estadística se refiere al análisis de las relaciones que se dan entre dos variables aleatorias. El análisis de estas relaciones está muy frecuentemente ligado al análisis de una variable, llamada **variable dependiente** (Y) , y del efecto que sobre ella tiene otra (u otras) variable(s), llamada(s) **variable(s) independiente(s)** (X), y permite responder a dos cuestiones básicas:

- ¿Es significativa la influencia que tiene la variable independiente sobre la variable dependiente?
- Si, en efecto, esa relación es significativa, ¿cómo es? y ¿podemos aprovechar esa relación para predecir valores de la variable dependiente a partir de valores observados de la variable independiente? Más aún, ¿podemos inferir características sobre esa relación y con el fenómeno que subyace a ella?

Ejemplo. Un equipo de investigadores que trabajan en seguridad en el trabajo está tratando de analizar cómo la piel absorbe un cierto componente químico peligroso. Para ello, coloca diferentes volúmenes del compuesto químico sobre diferentes segmentos de piel durante distintos intervalos de tiempo, midiendo al cabo de ese tiempo el porcentaje de volumen absorbido del compuesto. El diseño del experimento se ha

realizado para que la interacción esperable entre el tiempo y el volumen no influya sobre los resultados. Los datos aparecen en el Cuadro 10.1

Lo que los investigadores se cuestionan es si la cantidad de compuesto por un lado y el tiempo de exposición al que se somete por otro, influyen en el porcentaje que se absorbe. De ser así, sería interesante estimar el porcentaje de absorción de personas que se sometan a una exposición de una determinada cantidad, por ejemplo, durante 8 horas.

En una primera aproximación al problema, podemos observar una representación gráfica de los datos en los diagramas de dispersión o nubes de puntos de la Figura 10.1. ¿Qué afirmaríamos? Parece que sí hay una relación lineal más o menos clara (pero no definitiva) entre el tiempo de exposición y el porcentaje de absorción, pero ¿la hay entre el volumen y el porcentaje de absorción?

Experimento	Volumen	Tiempo	Porcentaje Absorbido
1	0.05	2	50.88
2	0.05	10	49.96
3	0.05	24	83.66
4	2.00	2	54.09
5	2.00	10	68.27
6	2.00	24	85.65
7	5.00	2	48.39
8	5.00	10	64.88
9	5.00	24	88.01

Cuadro 10.1: Datos sobre el experimento de la absorción del compuesto

Un modelo de **regresión lineal simple** para una variable, Y (**variable dependiente**), dada otra variable, X (**variable independiente**), es un modelo matemático que permite obtener una fórmula capaz de relacionar Y con X basada sólo en relaciones lineales, del tipo

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

En esta expresión:

- Y representa a la variable dependiente, es decir, a aquella variable que deseamos estudiar en relación con otras.
- X representa a la variable independiente, es decir, aquellas que creemos que puede afectar en alguna medida a la variable dependiente. La estamos notando en mayúscula, indicando que podría ser una variable aleatoria, pero habitualmente se considera que es una constante que el investigador puede fijar a su antojo en distintos valores.
- ε representa el **error aleatorio**, es decir, aquella cantidad (aleatoria) que provoca que la relación entre la variable dependiente y la variable independiente no sea perfecta, sino que esté sujeta a incertidumbre.

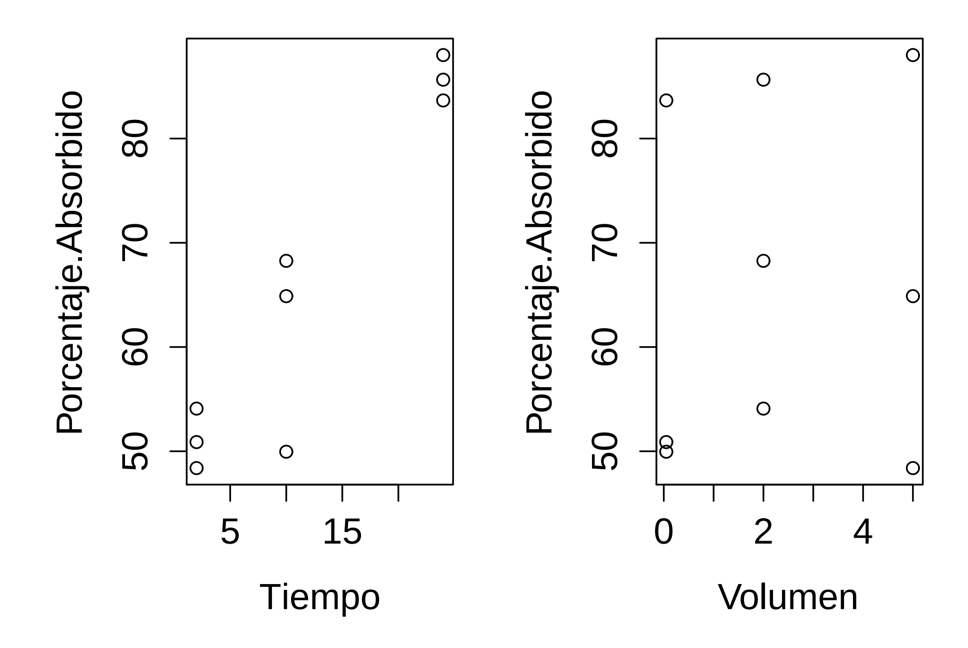


Figura 10.1: Nube de puntos

Hay que tener en cuenta que el valor de ε será siempre desconocido hasta que se observen los valores de X e Y, de manera que el modelo de predicción será realmente

$$\hat{Y} = \beta_0 + \beta_1 X.$$

Lo que en primer lugar resultaría deseable de un modelo de regresión es que estos errores aleatorios ocurran en la misma medida por exceso que por defecto, sea cual sea el valor de X, de manera que $E\left[\varepsilon/_{X=x}\right] = E\left[\varepsilon\right] = 0$ y, por tanto,

$$\begin{aligned} E\left[Y/_{X=x}\right] &= \beta_0 + \beta_1 x + E\left[\varepsilon/_{X=x}\right] \\ &= \beta_0 + \beta_1 x. \end{aligned}$$

Es decir, las medias de los valores de Y para un valor de X dado son una recta.

La Figura 10.2 representa una nube de puntos y la recta de regresión que los ajusta de unos datos genéricos. Podemos ver el valor concreto de $\varepsilon = y - E\left[Y/_{X=x}\right]$ para un dato, supuesto que hemos obtenido un modelo de regresión. En ella se puede ver también la interpretación de los coeficientes del modelo:

- β_0 es **la ordenada al origen** del modelo, es decir, el punto donde la recta intercepta o corta al eje y.

- β_1 representa **la pendiente** de la línea y, por tanto, puede interpretarse como el incremento de la variable dependiente por cada incremento en una unidad de la variable independiente.

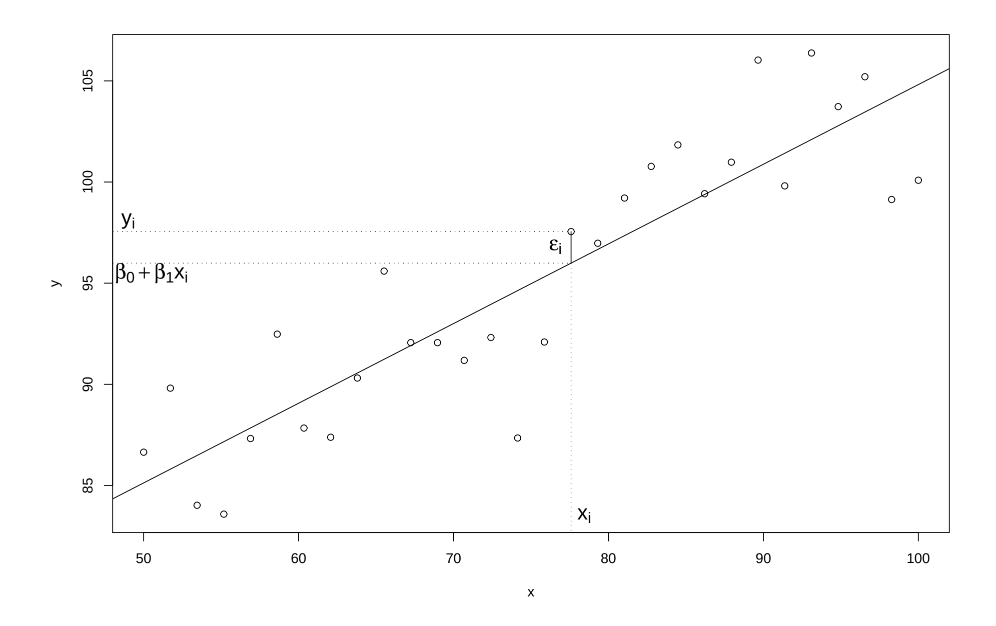


Figura 10.2: Diagrama de dispersión y línea de las medias hipotéticas.

Nota. Es evidente que la utilidad de un modelo de regresión lineal tiene sentido siempre que la relación hipotética entre X e Y sea de tipo lineal, pero ¿qué ocurre si en vez de ser de este tipo es de otro tipo (exponencial, logarítmico, hiperbólico...)?

En primer lugar, es absolutamente conveniente dibujar el diagrama de dispersión antes de comenzar a tratar de obtener un modelo de regresión lineal, ya que si la forma de este diagrama sugiere un perfil distinto al de una recta quizá deberíamos plantearnos otro tipo de modelo.

Y, por otra parte, si se observa que el diagrama de dispersión es de otro tipo conocido, puede optarse por realizar un cambio de variable para considerar un modelo lineal. Existen técnicas muy sencillas para esta cuestión, pero no las veremos aquí.

10.2. Estimación de los coeficientes del modelo por mínimos cuadrados

Si queremos obtener el modelo de regresión lineal *que mejor se ajuste a los datos de la muestra*, deberemos estimar los coeficientes β_0 y β_1 del modelo. Para obtener estimadores de estos coeficientes vamos a considerar un nuevo método de estimación, conocido como **método de mínimos cuadrados**. Hay que decir que bajo determinados supuestos que veremos en breve, los estimadores de mínimos cuadrados coinciden con los estimadores máximo-verosímiles de β_0 y β_1.

El razonamiento que motiva el método de mínimos cuadrados es el siguiente: si tenemos una muestra de

valores de las variables independiente y dependiente,

$$(x_1, y_1), ..., (x_n, y_n),$$

buscaremos valores estimados de β_0 y β_1, que notaremos por $\hat{\beta}_0$ y $\hat{\beta}_1$, de manera que en el modelo ajustado,

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x$$

minimice la suma de los cuadrados de los errores observados. Recordemos que

$$E\left[Y/_{X=x}\right] = \beta_0 + \beta_1 x,$$

luego $\hat{y}_x$ puede interpretarse de dos formas:

1. Como una predicción del valor que tomará Y si $X = x$.

2. Como una estimación del valor medio de Y cuando $X = x$.

Concretando, lo que buscamos es minimizar la **suma de los cuadrados de los errores**

$$SSE = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2,$$

es decir buscamos

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\left[\min_{\beta_0,\beta_1} SSE\right].$$

Se llama **recta de regresión por mínimos cuadrados (o simplemente recta de regresión) de** Y **dada** X a la línea que tiene la SSE más pequeña de entre todos los modelos lineales.

La solución de ese problema de mínimo se obtiene por el mecanismo habitual: se deriva SSE respecto de $\hat{\beta}_0$ y $\hat{\beta}_1$, se iguala a cero y se despejan estos. La solución es $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$ y $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, donde

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$
$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

Con esta notación, es fácil demostrar que

$$SSE = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2 = \frac{SS_{xx}SS_{yy} - SS_{xy}^2}{SS_{xx}}$$
$$= SS_{yy} - \frac{SS_{xy}2}{SS_{xx}} = SS_{yy} - SS_{xy} \times \hat{\beta}_1.$$

En este sentido, se define como medida de la calidad del ajuste de la recta de regresión el *error estandar del ajuste* como

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_i \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x\right)\right)^2}{n-2}}$$
$$= \sqrt{\frac{SS_{yy} - \hat{\beta}_1 SS_{xy}}{n-2}}.$$

Cuanto mayor sea esta cantidad, peor son las predicciones de la recta de regresión.

Ejemplo. Para los datos sobre el ejemplo de la absorción del compuesto, vamos a calcular e interpretar las dos rectas de regresión posibles.

En primer lugar, vamos a considerar la recta de regresión para explicar el porcentaje de absorción (y) conocido el volumen de sustancia (x):

$$SS_{xy} = 36.24, \; SS_x = 37.31$$

luego

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = 0.97$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 63.69,$$

así que la recta de regresión ajustada es

$$\hat{y}_x = 63.69 + 0.97 \times x.$$

La interpretación de $\hat{\beta}_1 = 0.97$ es que el porcentaje de absorción, Y, aumenta en promedio 0.97 por cada incremento de 1 unidad de volumen de compuesto. La interpretación de $\hat{\beta}_0 = 63.69$ sería la del valor promedio de Y cuando $x = 0$, pero es que en este caso este supuesto no tiene sentido, así que no debe tenerse en cuenta.

Vamos con la recta de regresión para explicar el porcentaje de absorción (y) en función del tiempo de exposición (x):

$$SS_{xy} = 1187.96, \; SS_{xx} = 744$$

luego

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = 1.60$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 46.82,$$

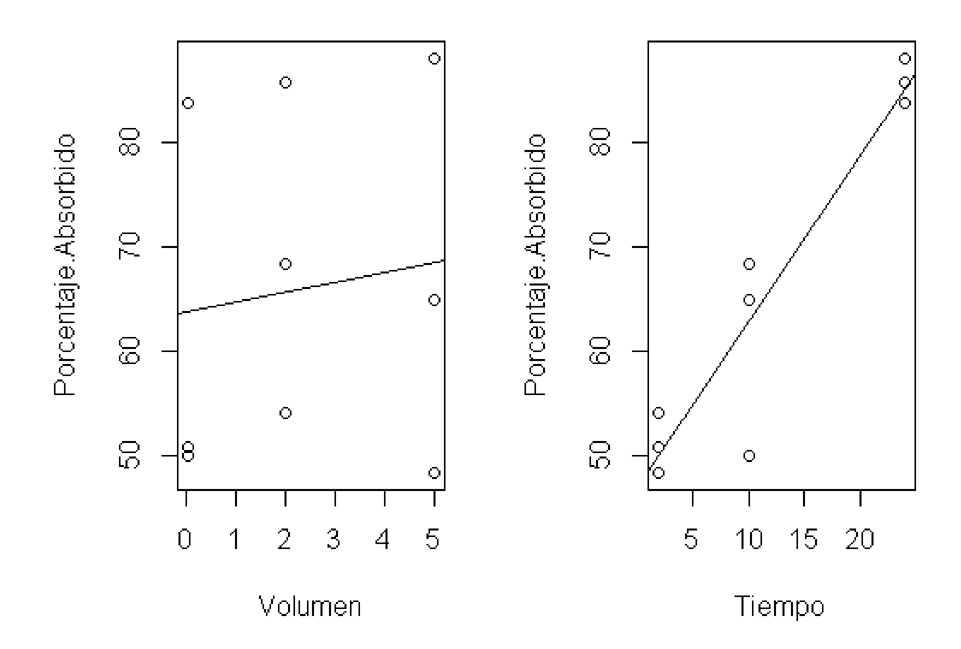


Figura 10.3: Nubes de puntos y rectas de regresión ajustadas en el ejemplo de la absorción

así que la recta de regresión ajustada es

$$\hat{y}_x = 46.82 + 1.60 \times x.$$

Por cada incremento de una unidad del tiempo de exposición, el porcentaje de absorción aumenta en media 1.60.

Ahora vamos a representar las nubes de puntos de nuevo con sus rectas de regresión ajustadas. De esa manera podremos comprobar de una forma gráfica cómo de buenas son las rectas en cuanto a su capacidad de ajuste de los datos. Los resultados aparecen en la Figura 10.3. Podemos ver que el ajuste es mucho mejor cuando la variable explicativa es el tiempo de absorción, mientras que si la variable explicativa es el volumen, la recta no puede pasar cerca de los datos.

Nota. Hay que hacer una observación importante que suele conducir a frecuentes errores. La recta de regresión para la variable dependiente Y, dada la variable independiente X no es la misma que la recta de regresión de X dada Y. La razón es muy sencilla: para obtener la recta de regresión de Y dado X debemos minimizar

$$\sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right)^2,$$

mientras que para obtener la recta de regresión de X dado Y deberíamos minimizar

$$\sum_{i=1}^{n}\left(x_i - \left(\hat{\beta}_0 + \hat{\beta}_1 y_i\right)\right)^2,$$

en cuyo caso obtendríamos como solución

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{yy}}$$
$$\hat{\beta}_0 = \bar{x} - \hat{\beta}_1 \bar{y},$$

siendo la recta de regresión, $\hat{x} = \hat{\beta}_0 + \hat{\beta}_1 y$.

El error que suele cometerse con frecuencia es pensar que si tenemos, por ejemplo, la recta de Y dado X, la de X dado Y puede obtenerse *despejando*.

Es importante que, para terminar este apartado, recordemos que $\hat{\beta}_0$ y $\hat{\beta}_1$ son sólo estimaciones de β_0 y β_1, estimaciones basadas en los datos que se han obtenido en la muestra.

Una forma de hacernos conscientes de que se trata de estimaciones y no de valores exactos (es imposible conocer el valor exacto de ningún parámetro poblacional) es proporcionar las estimaciones de los errores estandar de las estimaciones de β_0 y β_1. Se conoce que dichas estimaciones son:

$$s.e.\left(\hat{\beta}_1\right) = \sqrt{\frac{s_e^2}{SS_{xx}}}$$
$$s.e.\left(\hat{\beta}_0\right) = \sqrt{s_e^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}\right)}$$

Ejemplo. En el ejemplo de los datos de absorción hemos estimado los coeficientes de las dos rectas de regresión del porcentaje de absorción en función del volumen y del tiempo de absorción. Vamos a completar ese análisis con el cálculo de los errores estandares de esas estimaciones. Los resultados aparecen resumidos en la siguiente tabla:

Modelo	$\hat{\beta}_0$	$s.e.\left(\hat{\beta}_0\right)$	$\hat{\beta}_1$	$s.e.\left(\hat{\beta}_1\right)$
$\%\ absorción = \beta_0 + \beta_1 \times Volumen$	63.69	8.80	0.97	2.83
$\%\ absorción = \beta_0 + \beta_1 \times Tiempo$	46.82	3.16	1.60	0.21

Obsérvese que los errores estandar en el modelo en función del volumen son mayores proporcionalmente que en el modelo en función del tiempo de absorción.

10.3. Supuestos adicionales para los estimadores de mínimos cuadrados

Hasta ahora lo único que le hemos exigido a la recta de regresión es:

1. Que las medias de Y para cada valor de x se ajusten *más o menos* a una línea recta, algo fácilmente comprobable con una nube de puntos. Si el aspecto de esta nube no recuerda a una línea recta sino a otro tipo de función, lógicamente no haremos regresión lineal.

2. Que los errores tengan media cero, independientemente del valor de x, lo que, por otra parte, no es una hipótesis sino más bien un requerimiento lógico al modelo.

Lo que ahora vamos a hacer es añadir algunos supuestos al modelo de manera que cuando éstos se cumplan, las propiedades de los estimadores de los coeficientes del modelo sean muy buenas. Esto nos va a permitir hacer inferencia sobre estos coeficientes y sobre las estimaciones que pueden darse de los valores de la variable dependiente.

Los supuestos que podemos añadir se refieren al error del modelo, la variable ε.

Supuesto 1. Tal y como ya hemos dicho, $E\left[\epsilon/_{X=x}\right] = E\left[\epsilon\right] = 0$, lo que implica que $E\left[Y/_{X=x}\right] = \beta_0 + \beta_1 x$.

Supuesto 2. La varianza de ϵ también es constante para cualquier valor de x dado, es decir, $Var\left(\epsilon/_{X=x}\right) = \sigma^2$ para todo x.

Supuesto 3. La distribución de probabilidad de ϵ es normal.

Supuesto 4. Los errores ϵ son independientes unos de otros, es decir, la magnitud de un error no influye en absoluto en la magnitud de otros errores.

En resumen, todos los supuestos pueden resumirse diciendo que $\epsilon\mid_{X=x} \to N(0,\sigma^2)$ y son independientes entre sí.

Estos supuestos son restrictivos, por lo que deben comprobarse cuando se aplica la técnica. Si el tamaño de la muestra es grande, la hipótesis de normalidad de los residuos estará bastante garantizada por el teorema central del límite. En cuanto a la varianza constante respecto a los valores de x, un incumplimiento moderado no es grave, pero sí si las diferencias son evidentes.

Existen técnicas específicas para evaluar en qué medida se cumplen estas hipótesis. También existen procedimientos para corregir el incumplimiento de estos supuestos. Estos aspectos serán tratados al final del tema.

10.4. Inferencias sobre el modelo

10.4.1. Inferencia sobre la pendiente

Al comienzo del capítulo nos planteábamos como uno de los objetivos de la regresión el decidir si el efecto de la variable independiente es o no significativo para la variable dependiente. Si nos fijamos, esto es equivalente a contrastar si el coeficiente β_1 es o no significativamente distinto de cero. Vamos a profundizar en porqué es así.

Observemos la Figura 10.4. En la nube de puntos y la recta de regresión ajustada de la izquierda, ¿observamos una relación lineal *buena* entre x e y con un buen ajuste de la recta de regresión? Cabría pensar que sí, pero

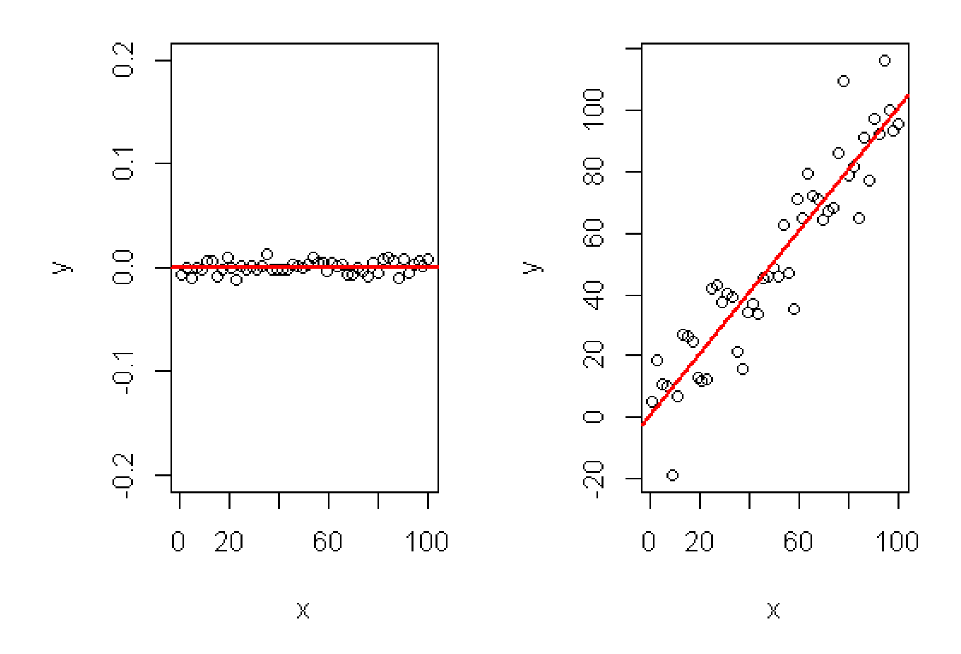


Figura 10.4: Nubes de puntos y rectas de regresión que las ajustan

estaríamos equivocados: si la recta de regresión trata de explicar y en función de x, ¿cuánto varía y conforme varía x? Dado que la pendiente de esa recta es cero o prácticamente cero, por mucho que cambies x, eso no afecta al valor de y, es decir, ¡x **no influye nada sobre** y**!** Sin embargo, en la nube de puntos de la derecha, a pesar de que aparentemente el ajuste es peor, la recta ajustada sí tiene pendiente distinta de cero, luego el hecho de que y varíe viene dado en buena parte por el hecho de que x varía, y ello ocurre porque la pendiente de esa recta es distinta de cero. Así pues, no lo olvidemos: decir que dos variables están relacionadas linealmente equivale a decir que la pendiente de la recta de regresión que ajusta una en función de la otra es distinta de cero.

Pues bien, dados los supuestos descritos en la sección anterior, es posible obtener un contraste de este tipo, tal y como se resumen en el Cuadro 10.2. En ella, si, en efecto, lo que deseamos es contrastar si el efecto de la variable independiente es o no significativo para la variable dependiente, el valor de b_1 será cero.

Ejemplo. Para los datos del ejemplo sobre la absorción, partíamos del deseo de comprobar si al volumen y/o el tiempo de exposición influían sobre el porcentaje de absorción. Las nubes de puntos y el ajuste de la recta ya nos dieron pistas: daba la impresión de que el tiempo de absorción sí influía en el porcentaje de absorción, pero no quedaba tan claro si el volumen lo hacía. Es el momento de comprobarlo.

Nos planteamos en primer lugar si el tiempo de exposición influye o no sobre el porcentaje de absorción, es decir, nos planteamos si en el modelo lineal

$$Porcentaje\ de\ absorción = \beta_0 + \beta_1 \times Tiempo\ de\ exposición + \varepsilon$$

Tipo de prueba	Unilateral a la izquierda	Bilateral	Unilateral a la derecha
Hipótesis	$H_0 : \beta_1 = b_1$ $H_1 : \beta_1 < b_1$	$H_0 : \beta_1 = b_1$ $H_1 : \beta_1 \neq b_1$	$H_0 : \beta_1 = b_1$ $H_1 : \beta_1 > b_1$
Estadístico de contraste	$t = \frac{\hat{\beta}_1 - b_1}{\sqrt{s_e^2/SS_{xx}}}$, $s_e^2 = \frac{SS_{yy} - \hat{\beta}_1 SS_{xy}}{n-2} = \frac{SSE}{n-2}$		
Región de rechazo	$t < t_{\alpha;n-2}$	$\|t\| > t_{1-\alpha/2;n-2}$	$t > t_{1-\alpha;n-2}$
p-valor	$P[T_{n-2} < t]$	$2P[T_{n-2} > \|t\|]$	$P[T > t]$
Supuestos	Los dados en la Sección 10.3		

Cuadro 10.2: Contraste sobre β_1

el coeficiente β_1 es o no cero. Formalmente, nos planteamos $H_0 : \beta_1 = 0$ frente a $H_1 : \beta_1 \neq 0$:

$$\hat{\beta}_1 = 1.6$$
$$s_e^2 = \frac{SS_{yy} - \hat{\beta}_1 SS_{xy}}{n-2} = 32.82$$
$$t_{0.975;9-2} = 2.364624,\ t_{0.025;30-2} = -2.364624$$
$$t = \frac{1.6 - 0}{\sqrt{32.82/744}} = 7.60,$$

luego, como cabía esperar, podemos afirmar a la luz de los datos y con un 95 % de confianza que el efecto del tiempo de exposición sobre el porcentaje de absorción es significativo. El p-valor, de hecho, es $p = 2P[T_7 > 7.60] = 0.000126$.

Vamos ahora a analizar si el efecto lineal del volumen sobre el porcentaje de absorción es significativo. Es decir, ahora nos planteamos si en el modelo lineal

$$Porcentaje\ de\ absorción = \beta_0 + \beta_1 \times Volumen + \varepsilon$$

el coeficiente β_1 es o no cero, es decir, planteamos el contraste de $H_0 : \beta_1 = 0$ frente a $H_1 : \beta_1 \neq 0$:

$$\hat{\beta}_1 = 0.97$$
$$s_e^2 = \frac{SS_{yy} - \hat{\beta}_1 SS_{xy}}{n-2} = 298.77$$
$$t_{0.975;9-2} = 2.364624,\ t_{0.025;30-2} = -2.364624$$
$$t = \frac{0.97 - 0}{\sqrt{298.77/37.31}} = 0.34,$$

luego, como cabía esperar, no podemos afirmar a la luz de los datos y con un 95 % de confianza que el efecto del volumen sobre el porcentaje de absorción sea significativo. El p-valor, de hecho, es $p = 2P[T_7 > 0.34] = 0.741$.

En vista de los resultados, a partir de ahora dejaremos de considerar el efecto del volumen sobre el porcentaje de absorción, y sólo tendremos en cuenta el efecto del tiempo de exposición.

Ejemplo. Un ingeniero químico está calibrando un espectrómetro para medir la concentración de CO en muestras de aire. Esta calibración implica que debe comprobar que no hay diferencias *significativas* entre la concentración verdadera de CO (x) y la concentración medida por el espectrómetro (y). Para ello toma 11 muestras de aire en las que conoce su verdadera concentración de CO y las compara con la concentración medida por el espectrómetro. Los datos son los siguientes (las unidades son ppm):

x	0	10	20	30	40	50	60	70	80	90	100
y	1	12	20	29	38	48	61	68	79	91	97

Lo ideal, lo deseado, sería que $y = x$, es decir, que el modelo lineal que explica y en función de x tuviera coeficientes $\beta_0 = 0$ y $\beta_1 = 1$. Por ahora vamos a centrarnos en el primer paso en la comprobación de que el espectrómetro está bien calibrado, que implica contrastar que $\beta_1 = 1$. Para ello,

$$SS_{xx} = 11000;\ SS_{yy} = 10506.73;\ SS_{xy} = 10740$$
$$\hat{\beta}_1 = \frac{10460}{11000} = 0.976$$
$$s_e^2 = \frac{SS_{yy} - \hat{\beta}_1 SS_{xy}}{n-2} = 2.286$$

por lo tanto,

$$t = \frac{0.976 - 1}{\sqrt{1.964/11000}} = -1.639.$$

Dado que $t_{1-\frac{0.05}{2};11-2} = t_{0.975;9} = 2.262$ y $|-1.639| < 2.262$, no hay razones para concluir que $\beta_1 \neq 1$. Así pues, el modelo podría ser

$$y = \beta_0 + x,$$

aunque lo deseado, insistamos, sería que fuera

$$y = x,$$

es decir, que lo que mida el espectrómetro coincida con la cantidad real de CO en el aire. Como hemos dicho, eso ocurriría si $\beta_0 = 0$, lo que equivale a decir que en ausencia de CO, el espectrómetro esté a cero.

Además del contraste de hipótesis, es trivial proporcionar un intervalo de confianza para la pendiente, ya que conocemos su estimación, su error estandar y la distribución en el muestreo (t-student, como aparece en el contraste). Concretamente,

$$P\left[\beta_1 \in \left(\hat{\beta}_1 - t_{1-\frac{\alpha}{2};n-2} \times s.e.\left(\hat{\beta}_1\right), \hat{\beta}_1 + t_{1-\frac{\alpha}{2};n-2} \times s.e.\left(\hat{\beta}_1\right)\right)\right] = 1 - \alpha.$$

Ejemplo. En el ejemplo que acabamos de ver sobre la calibración del espectrómetro, el intervalo de confianza para β_1 es $(0.94,\ 1.01)$. Como podemos ver, el valor $\beta_1 = 1$ es un valor confiable del intervalo, luego ratificamos que no podemos afirmar que el espectrómetro esté mal calibrado.

Tipo de prueba	Unilateral a la izquierda	Bilateral	Unilateral a la derecha
Hipótesis	$H_0 : \beta_0 = b_0$ $H_1 : \beta_0 < b_0$	$H_0 : \beta_0 = b_0$ $H_1 : \beta_0 \neq b_0$	$H_0 : \beta_0 = b_0$ $H_1 : \beta_0 > b_0$
Estadístico de contraste	$t = \frac{\hat{\beta}_0 - b_0}{\sqrt{s_e^2\left(\frac{1}{n}+\frac{\bar{x}^2}{SS_{xx}}\right)}}$, $s_e^2 = \frac{SS_{yy}-\hat{\beta}_1 SS_{xy}}{n-2} = \frac{SSE}{n-2}$		
Región de rechazo	$t < t_{\alpha;n-2}$	$\|t\| > t_{1-\alpha/2;n-2}$	$t > t_{1-\alpha;n-2}$
p-valor	$P[T_{n-2} < t]$	$2P[T_{n-2} > \|t\|]$	$P[T > t]$
Supuestos	Los dados en la Sección 10.3		

Cuadro 10.3: Contraste sobre β_0

10.4.2. Inferencia sobre la ordenada en el origen

Este último ejemplo pone de manifiesto que también puede tener interés realizar contrastes sobre el valor de β_0. Para ello, el Cuadro 10.3 describe el procedimiento de un contraste de este tipo.

Finalmente, tengamos en cuenta que podría ser de interés un contraste conjunto sobre β_0 y β_1, por ejemplo, del tipo $\beta_0 = 0$, $\beta_1 = 1$. Hay que decir que este tipo de contrastes múltiples superan los contenidos de esta asignatura. Lo único que podríamos hacer en un contexto como el nuestro es realizar sendos contrastes sobre β_0 y β_1 por separado, teniendo en cuenta el nivel de significación de ambos contrastes.

Ejemplo. En el ejemplo anterior, vamos a contrastar si, en efecto, $\beta_0 = 0$, lo que equivaldrá a concluir que no hay razones para pensar que el espectrómetro está mal calibrado. Para ello,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 0.636$$

por lo tanto,

$$t = \frac{0.636 - 0}{\sqrt{2.286\left(\frac{1}{11} + \frac{50^2}{11000}\right)}} = 0.746.$$

Comoquiera que $0.746 < t_{0.975;9} = 2.261$, tampoco tenemos razones para pensar que $\beta_0 = 0$ con un 95 % de confianza, luego, en resumen, no existen razones para pensar que el espectrómetro está mal calibrado.

Ejemplo. Imaginemos que deseamos comprobar experimentalmente que, tal y como predice la ley de Ohm, la tensión (V) entre los extremos de una resistencia y la intensidad de corriente (I) que circula por ella se relacionan siguiendo la ley

$$V = R \times I,$$

donde R es el valor de la resistencia. Nosotros vamos a realizar la comprobación con una misma resistencia, variando los valores de la intensidad, por lo que la ecuación equivale a

$$V = \beta_0 + \beta_1 \times I,$$

siendo $\beta_0 = 0$ y $\beta_1 = R$. Los datos son los que aparecen en el Cuadro 10.4.

Tenemos que realizar un contraste, $H_0 : \beta_0 = 0$ frente a $H_1 : \beta_0 \neq 0$ que equivale a contrastar en realidad

Observación	I (mA)	V (V)
1	0.16	0.26
2	6.54	1.04
3	12.76	2.02
4	19.26	3.05
5	25.63	4.06
6	31.81	5.03
7	38.21	6.03
8	47.40	7.03
9	54.00	8.06
10	60.80	8.99
11	68.00	10.01

Cuadro 10.4: Datos para la comprobación de la Ley de Ohm

que nuestros aparatos de medida están bien calibrados, puesto que la ley de Ohm obliga a que $\beta_0 = 0$. Vamos allá:

$$\begin{aligned} SS_{xx} &= 5105.90 \\ SS_{yy} &= 107.25 \\ SS_{xy} &= 739.49 \\ \hat{\beta}_1 &= 0.14 \\ \hat{\beta}_0 &= 0.25 \\ s_e^2 &= 0.022 \end{aligned}$$

Así pues,

$$t = \frac{0.25 - 0}{\sqrt{0.022\left(\frac{1}{11} + \frac{33.14^2}{5105.90}\right)}} = 3.531.$$

Dado que $t_{0.975,9} = 2.262$, tenemos que rechazar la hipótesis $H_0 : \beta_0 = 0$, lo que **¡contradice la ley de Ohm!** Lo que este análisis pone de manifiesto es que tenemos algún problema en nuestras mediciones.

Dejemos un poco de lado este último resultado. Si queremos estimar el valor de la resistencia, una estimación puntual es, como hemos visto, $\hat{R} = \hat{\beta}_1 = 0.14$, y un intervalo de confianza al 95 % de confianza (omitimos los detalles de los cálculos) resulta ser $(0.141, 0.149)$.

Finalmente, podemos también proporcionar un intervalo de confianza para la ordenada en el origen, dado por

$$P\left[\beta_0 \in \left(\hat{\beta}_0 - t_{1-\frac{\alpha}{2};n-2} \times s.e.\left(\hat{\beta}_0\right), \hat{\beta}_0 + t_{1-\frac{\alpha}{2};n-2} \times s.e.\left(\hat{\beta}_0\right)\right)\right] = 1 - \alpha.$$

Ejemplo. En el ejemplo del espectrómetro, el intervalo de confianza para la ordenada en el origen es $(-1.29, 2.57)$, luego es confiable pensar que $\beta_0 = 0$. En suma, hemos comprobado que es posible $\beta_1 = 1$ y $\beta_0 = 0$, luego hemos comprobado que la ecuación $y = x$ no puede ser rechazada con los datos disponibles, es decir, que no hay razones para pensar que el espectrómetro esté mal calibrado.

Ejemplo. En el ejemplo de la comprobación de la Ley de Ohm, el intervalo de confianza al 95 % para la ordenada en el origen es $(0.09, 0.41)$. Dado que ese intervalo no incluye al cero, podemos afirmar con un 95 % de confianza que la recta de regresión no pasa por el origen, lo que contradice la Ley de Ohm.

10.5. El coeficiente de correlación lineal

$\hat{\beta}_1$ mide en cierto modo la relación que existe entre la variable dependiente y la variable independiente, ya que se interpreta como el incremento que sufre Y por cada incremento unitario de X. Sin embargo, es una medida sujeta a la escala de las variables X e Y, de manera que se hace difícil poder comparar distintos $\hat{\beta}_1' s$ entre sí.

En esta sección vamos a definir el llamado **coeficiente de correlación lineal**, que ofrece una medida cuantitativa de la fortaleza de la relación lineal entre X e Y en la muestra, pero que a diferencia de $\hat{\beta}_1$, es adimensional, ya que sus valores siempre están entre -1 y 1, sean cuales sean las unidades de medida de las variables.

Dada una muestra de valores de dos variables $(x_1, y_1), ..., (x_n, y_n)$, el **coeficiente de correlación lineal muestral** r se define como

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{\sqrt{SS_{xx}}}{\sqrt{SS_{yy}}}\hat{\beta}_1.$$

Como comentábamos, la interpretación del valor de r es la siguiente:

- r cercano o igual a 0 implica poca o ninguna relación lineal entre X e Y.
- Cuanto más se acerque a 1 ó -1, más fuerte será la relación lineal entre X e Y.
- Si $r = \pm 1$, todos los puntos caerán exactamente en la recta de regresión.
- Un valor positivo de r implica que Y tiende a aumentar cuando X aumenta, y esa tendencia es más acusada cuanto más cercano está r de 1.
- Un valor negativo de r implica que Y disminuye cuando X aumenta, y esa tendencia es más acusada cuanto más cercano está r de -1.

Nota. En la Figura 10.5 aparecen algunos de los supuestos que acabamos de enunciar respecto a los distintos valores de r. Hay que hacer hincapié en que r sólo es capaz de descubrir la presencia de relación de tipo lineal. Si, como en el último gráfico a la derecha de esta figura, la relación entre X e Y no es de tipo lineal, r no es adecuado como indicador de la fuerza de esa relación.

Nota. En la Figura 10.6 aparece un valor atípico entre un conjunto de datos con una relación lineal más que evidente. Por culpa de este dato, el coeficiente de correlación lineal será bajo. ¿Qué debe hacerse en

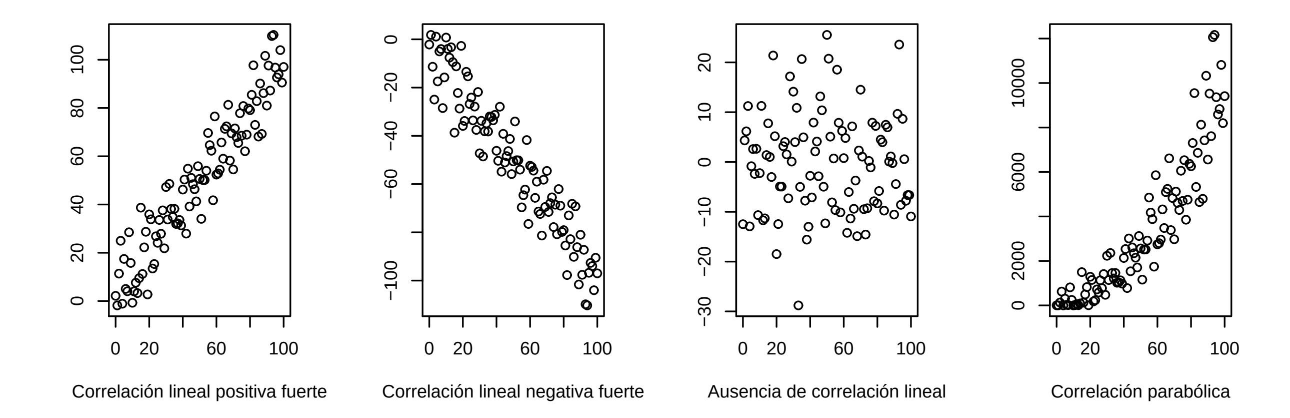


Figura 10.5: Valores de r y sus implicaciones.

este caso? En general, no se deben eliminar datos de una muestra, pero podría ocurrir que datos atípicos correspondan a errores en la toma de las muestras, en el registro de los datos o, incluso, que realmente no procedan de la misma población que el resto de los datos: en ese caso, eliminarlos podría estar justificado de cara a analizar de una forma más precisa la relación lineal entre los datos.

Nota. Correlación frente a causalidad. Hay que hacer una advertencia importante acerca de las interpretaciones del coeficiente de correlación lineal. Es muy frecuente que se utilice para justificar relaciones causa-efecto, y eso es un grave error. r sólo indica presencia de relación entre las variables, pero eso no permite inferir, por ejemplo, que un incremento de X sea la causa de un incremento o una disminución de Y.

Ejemplo. Para los datos del ejemplo sobre la absorción, calculemos r e interpretémoslo.

En el caso del porcentaje de absorción en función del volumen de compuesto,

$$r = \frac{36.24}{\sqrt{37.30 \times 2126.61}} = 0.129;$$

vemos que la relación es muy pequeña; de hecho, comprobamos mediante un contraste de hipótesis sobre β_1 que era no significativa.

En el caso del porcentaje de absorción en función del tiempo de absorción,

$$r = \frac{36.24}{\sqrt{744 \times 2126.61}} = 0.944.$$

Esta relación sí resulta ser muy fuerte y en sentido directo. Por eso al realizar el test sobre β_1, éste sí resultó ser significativo.

No podemos olvidar que el coeficiente de correlación lineal muestral, r, mide la correlación entre los valores

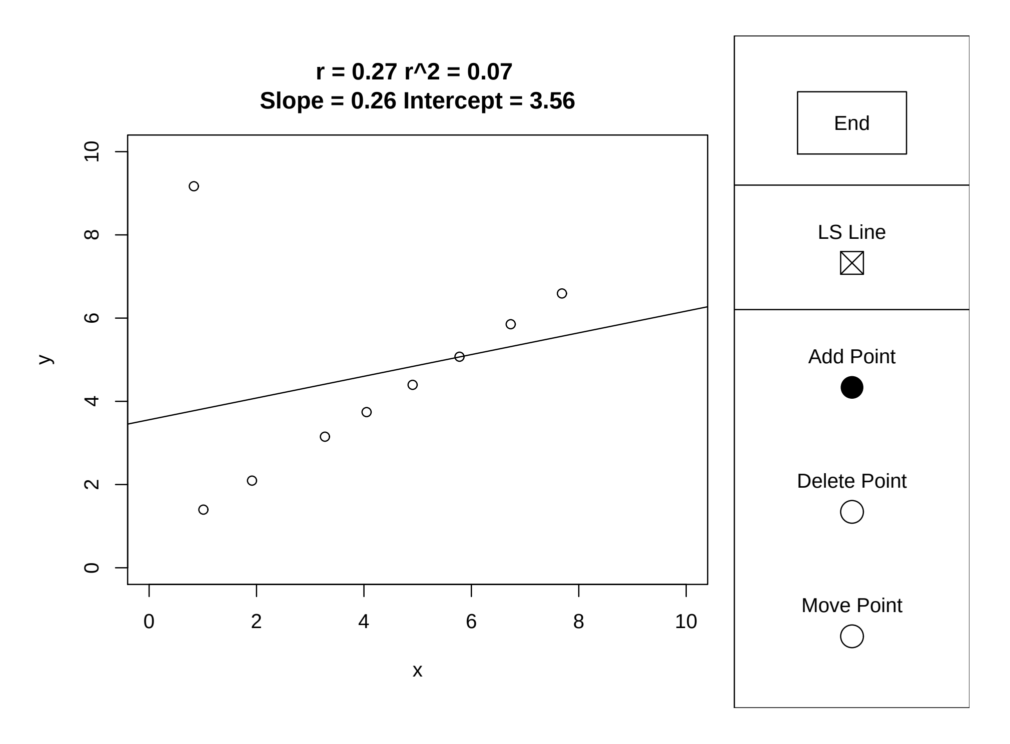


Figura 10.6: Un dato atípico entre datos relacionados linealmente.

de X y de Y en la muestra. Existe un coeficiente de correlación lineal similar pero que se refiere a todos los posibles valores de la variable. Evidentemente, r es un estimador de este coeficiente poblacional.

Dadas dos variables X e Y, el **coeficiente de correlación lineal poblacional,** ρ, se define como[a]

$$\rho = \frac{E\left[(X - EX)(Y - EY)\right]}{\sqrt{VarXVarY}} = \frac{\sqrt{VarX}}{\sqrt{VarY}}\beta_1.$$

[a]Este concepto se estudia también en el capítulo de vectores aleatorios.

Inmediatamente surge la cuestión de las inferencias. Podemos y debemos utilizar r para hacer inferencias sobre ρ. De todas formas, en realidad estas inferencias son equivalentes a las que hacemos sobre β_1, ya que la relación entre β_1 y ρ provoca que la hipótesis $H_0 : \beta_1 = 0$ sea equivalente a la hipótesis $H_0 : \rho = 0$. Podemos, por lo tanto, utilizar el contraste resumido en el Cuadro 10.2 para $b_1 = 0$ y teniendo en cuenta que

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Ejemplo. Vamos a contrastar $H_0 : \rho = 0$ frente a $H_1 : \rho \neq 0$ de nuevo en el ejemplo de la absorción. El estadístico de contraste es $t = \frac{0.944\times\sqrt{9-2}}{\sqrt{1-0.944^2}} = 7.60$, que coincide con el valor de t cuando contrastamos $H_0 : \beta_1 = 0$, frente a $H_1 : \beta_1 \neq 0$. Vemos que, en efecto, es el mismo contraste.

10.6. Fiabilidad de la recta de regresión. El coeficiente de determinación lineal

Como hemos visto, el coeficiente de correlación lineal puede interpretarse como una medida de la contribución de una variable a la predicción de la otra mediante la recta de regresión. En esta sección vamos a ver una medida más adecuada para valorar hasta qué punto la variable independiente contribuye a predecir la variable dependiente.

Recordemos lo que habíamos observado en la Figura 10.4. Allí teníamos una recta, la de la izquierda, que aparentemente era *buena*, mientras que la de la derecha aparentemente era *peor*. Sin embargo, ya dijimos que eso era inexacto. En realidad nosotros no deseamos comprobar exactamente si los puntos están o no en torno a la recta de regresión, sino en qué medida la recta de regresión explica Y en función de X.

Vamos a entrar en detalles. Necesitamos que la recta explique Y en función de X porque Y tiene datos que presentan una cierta variabilidad: ¿cuánta variabilidad? Cuando definimos la varianza, esa variabilidad la medimos como

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 ,$$

de tal manera que cuanto más varíen los datos de Y mayor será SS_{yy}. Por otra parte, cuando ajustamos por la recta de regresión $\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 \times x$, medimos el error que cometemos en el ajuste con

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_x)^2 .$$

Vamos a ponernos en las dos situaciones límite que pueden darse en cuanto a la precisión de una recta de regresión:

- Si X no tiene ningún tipo de relación lineal con Y, entonces $\rho = 0$, en cuyo caso $\beta_1 = \frac{\sqrt{VarY}}{\sqrt{VarX}}\rho = 0$ y la recta es simplemente

  $$\begin{aligned} \hat{y}_i &= \beta_0 + \beta_1 x_i \\ &= \bar{y}. \end{aligned}$$

  Es decir, si X no tiene ningún tipo de relación lineal con Y, entonces la mejor predicción que podemos dar por el método de mínimos cuadrados es la media. Además, en ese caso

  $$\begin{aligned} SSE &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = SS_{yy}, \end{aligned}$$

  es decir, SSE es el total de la variación de los valores de Y. Está claro que esta es la peor de las situaciones posibles de cara a la precisión.

- Si la relación lineal entre X e Y es total, entonces $\rho = 1$, en cuyo caso $\beta_1 = \frac{\sqrt{VarY}}{\sqrt{VarX}}$. Además, si la

relación lineal es total, $y = \hat{y}_x$, de manera que

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 0.$$

Esta, desde luego, es la mejor de las situaciones posibles.

La idea de la medida que vamos a utilizar es cuantificar en qué medida estamos más cerca o más lejos de estas dos situaciones. Dado que SSE, que es la medida del error de la recta de regresión, puede ir de 0 (mejor situación posible) a SS_{yy} (peor situación posible), tan sólo tenemos que relativizar en una escala cómoda una medida de este error.

Se define el **coeficiente de determinación lineal** como

$$r^2 = 1 - \frac{SSE}{SS_{yy}}.$$

Nótese que la notación es r al cuadrado, ya que, en efecto, en una regresión lineal simple coincide con el coeficiente de correlación lineal al cuadrado.

Por lo tanto, la interpretación de r^2 es la medida en que X contribuye a la explicación de Y en una escala de 0 a 1, donde el 0 indica que el error es el total de la variación de los valores de Y y el 1 es la precisión total, el error 0. La medida suele darse en porcentaje. Dicho de otra forma:

Aproximadamente $100 \times r^2\,\%$ de la variación total de los valores de Y respecto de su media pueden ser explicada mediante la recta de regresión de Y dada X.

Ejemplo. En el ejemplo de la absorción explicada por el tiempo de exposición, $r^2 = 0.892$, de manera que podemos decir que el 89 % de la variación total de los valores del porcentaje de absorción puede ser explicada mediante la recta de mínimos cuadrados dado el tiempo de exposición. Es evidente que es un porcentaje importante, que proporcionará predicciones relativamente fiables.

10.7. Predicción y estimación a partir del modelo

Recordemos que en el modelo ajustado de la recta de regresión,

$$\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x$$

y, por otro lado,

$$E\left[Y/_{X=x}\right] = \beta_0 + \beta_1 x,$$

luego $\hat{y}_x$ puede interpretarse de dos formas:

1. Como **predicción** del valor que tomará Y cuando $X = x$.

2. Como **estimación** del valor medio de Y para el valor $X = x$, es decir, de $E\left[Y/_{X=x}\right]$.

Ambas cantidades están sujetas a incertidumbre, que será tanto mayor cuanto más variabilidad tenga Y, y/o peor sea el ajuste mediante la recta de regresión.

Lo que vamos a ver en esta sección para concluir el tema es cómo establecer *regiones de confianza* para estas predicciones de los valores de Y y para las estimaciones de los valores medios de Y dados valores de X. Estos resultados requieren que se verifiquen los supuestos adicionales sobre los errores dados en la sección 10.3.

Podemos garantizar con un $(1-\alpha)\times 100\,\%$ de confianza que cuando $X = x$, el valor medio de Y se encuentra en el intervalo

$$\left(\hat{y}_x - t_{1-\alpha/2;n-2} \times s_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}}, \hat{y}_x + t_{1-\alpha/2;n-2} \times s_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}}\right),$$

es decir, podemos garantizar que

$$P\left[E[Y/_{X=x}] \in \left(\hat{y}_x \mp t_{1-\alpha/2;n-2} \times s_e\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}}\right)|_{X=x}\right] = 1-\alpha.$$

Asimismo, podemos garantizar con un $(1-\alpha)\times 100\,\%$ de confianza que cuando $X = x$, el valor Y se encuentra en el intervalo

$$\left(\hat{y}_x - t_{1-\alpha/2;n-2} \times s_e\sqrt{1+\frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}}, \hat{y}_x + t_{1-\alpha/2;n-2} \times s_e\sqrt{1+\frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}}\right),$$

es decir, podemos garantizar que

$$P\left[Y \in \left(\hat{y}_x \mp t_{1-\alpha/2;n-2} \times s_e\sqrt{1+\frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}}\right)|_{X=x}\right] = 1-\alpha$$

Nota. No debemos olvidar que los modelos de regresión que podemos estimar lo son a partir de los datos de una muestra de valores de X e Y. A partir de estos modelos podemos obtener, como acabamos de recordar, predicciones y estimaciones para valores dados de X. Dado que el modelo se basa precisamente en **esos valores de la muestra**, no es conveniente hacer predicciones y estimaciones para valores de X que se encuentren fuera del rango de valores de X en la muestra.

Ejemplo. En la Figura 10.7 aparece la recta de regresión para los datos del ejemplo sobre la absorción del compuesto junto con líneas que contienen los intervalos de confianza al 95 % para las predicciones y las estimaciones asociadas a los distintos valores de X.

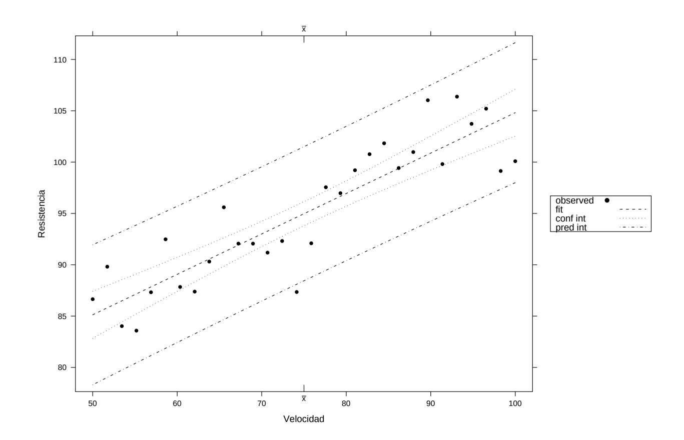


Figura 10.7: Recta de regresión con intervalos de confianza al 95 % para las predicciones (franjas más exteriores) y para las estimaciones (franjas interiores) en el ejemplo de la absorción.

Obsérvese que la amplitud de los intervalos se hace mayor en los valores más extremos de X. Es decir, los errores en las estimaciones y en las predicciones son mayores en estos valores más extremos. Esto debe ser un motivo a añadir al comentario anterior para no hacer estimaciones ni predicciones fuera del rango de valores de X en la muestra.

Por otra parte, nos planteábamos al comienzo de capítulo que sería de interés estimar el porcentaje de absorción que tendrá alguien que se someta a un tiempo de exposición al compuesto de 8 horas. Eso es una predicción, así que como estimación puntual daremos

$$\hat{y}_8 = 46.82 + 1.60 \times 8 = 59.59$$

y como intervalo de predicción al 95 %,

$$\left(\hat{y}_x \mp t_{1-\alpha/2;n-2} \times s_e \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}}\right) = \left(59.59 \mp 2.36 \times 5.73 \sqrt{1 + \frac{1}{9} + \frac{(8-12)^2}{744}}\right) = (45.17, 74.00) .$$

Por el contrario, imaginemos que los trabajadores de una empresa van a estar sometidos todos ellos a un tiempo de exposición de 8 horas. En ese caso, no tiene sentido que nos planteemos una predicción para saber cuál va a ser su porcentaje de absorción, ya que cada uno de ellos tendrá un porcentaje distinto; lo que sí tiene sentido es que nos planteemos cuál va a ser el porcentaje medio de absorción de los trabajadores sometidos a 8 horas de exposición al compuesto. Esto es un ejemplo de la estimación de un valor promedio. La estimación puntual es la misma que en la predicción, es decir, 59.59, pero el intervalo de confianza al 95 % es

$$\left(\hat{y}_x \mp t_{1-\alpha/2;n-2} \times s_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}}\right) = \left(59.59 \mp 2.36 \times 5.73 \sqrt{\frac{1}{9} + \frac{(8-12)^2}{744}}\right) = (54.66, 64.52) .$$

10.8. Diagnosis del modelo

Todo lo relacionado con inferencia sobre el modelo de regresión se ha basado en el cumplimiento de los supuestos descritos en el apartado 10.3. Como ya comentamos, en la medida en que todos o algunos de estos supuestos no se den, las conclusiones que se extraigan en la inferencia sobre el modelo podrían no ser válidas. Es por ello que es necesario comprobar estos supuestos mediante herramientas de diagnóstico. Aquí vamos a ver sólo las más básicas, vinculadas al análisis de los residuos y a la gráfica de residuos frente a los valores ajustados.

10.8.1. Normalidad de los residuos

Entre los supuestos del modelo consideramos que los residuos, es decir,

$$\epsilon_i = y_i - \hat{y}_i$$

siguen una distribución normal.

Ni que decir tiene que comprobar esta hipótesis en trivial: bastará con calcular los residuos, ajustarles una distribución normal y realizar un contraste de bondad de ajuste mediante, por ejemplo, el test de Kolmogorov-Smirnoff.

10.8.2. Gráfica de residuos frente a valores ajustados

El resto de supuestos se refieren a la varianza constante de los residuos, a su media cero y a su independencia.

Una de las herramientas diagnósticas más simples para estas hipótesis es la llamada *gráfica de residuos frente a valores ajustados*. Se trata de representar en unos ejes cartesianos:

1. En el eje X, los valores $\hat{y}_i$ de la muestra.
2. En el eje Y, los residuos, $\epsilon_i = y_i - \hat{y}_i$.

Habitualmente, se le añade a esta gráfica la recta de regresión de la nube de puntos resultante.

Vamos a ir viendo cómo debe ser esta gráfica en el caso de que se cumplan cada uno de los supuestos:

1. Si la media de los residuos es cero, la nube de puntos de la gráfica debe hacernos pensar en una recta de regresión horizontal situada en el cero, indicando que sea cual sea el valor $\hat{y}_i$, la media de los residuos es cero.
2. Si los errores son independientes, no debe observarse ningún *patrón* en la gráfica, es decir, ningún efecto en ella que haga pensar en algún tipo de relación entre $\hat{y}_i$ y ϵ_i.
3. Si los errores tienen una varianza constante (se habla entonces de **homocedasticidad**), la dispersión vertical de los puntos de la gráfica no debe variar según varíe el eje X. En caso contrario, se habla de **heterocedasticidad**.

Una última observación: si se dan todas las condiciones que acabamos de mencionar sobre la gráfica de residuos frente a valores ajustados, entonces es *probable*, pero no se tiene la seguridad, de que los supuestos del modelo sean ciertos.

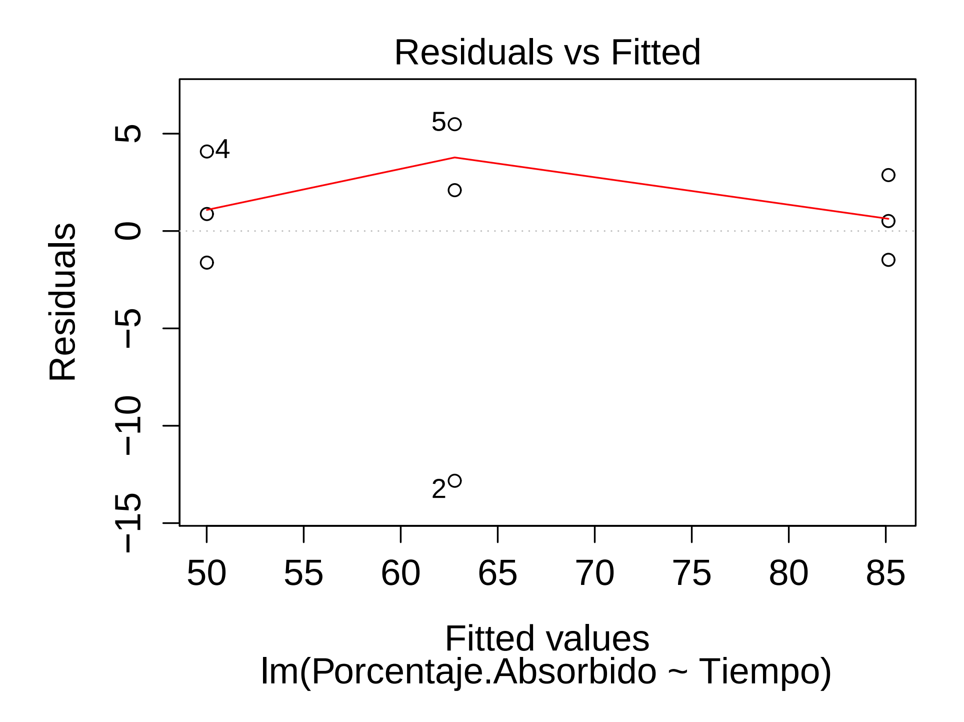


Figura 10.8: Gráfica de valores ajustados vs residuos en el ejemplo de la absorción

Ejemplo. Por última vez vamos a considerar el ejemplo de la absorción. En la Figura 10.8 aparece el gráfico de residuos vs valores ajustados y podemos ver que a primer vista parece que se dan las condiciones requeridas:

1. Los puntos se sitúan en torno al eje $Y = 0$, indicando que la media de los residuos parece ser cero.
2. No se observan patrones en los residuos.
3. No se observa mayor variabilidad en algunas partes del gráfico. Hay que tener en cuenta que son muy pocos datos para sacar conclusiones.

Parte IV

Procesos aleatorios

Capítulo 11

Procesos aleatorios

> The best material model of a cat is another, or preferably the same, cat.
>
> Norbert Wiener, *Philosophy of Science* (1945) (with A. Rosenblueth)

Resumen. Los procesos aleatorios suponen el último paso en la utilización de modelos matemáticos para describir fenómenos reales no determinísticos: concretamente, se trata de fenómenos aleatorios que dependen del tiempo. Se describen principalmente en términos de sus medias y sus covarianzas. En este capítulo se incluyen además algunos de los ejemplos más comunes de tipos de procesos y su comportamiento cuando se transmiten a través de sistemas lineales invariantes en el tiempo.

Palabras clave. Procesos aleatorios, función media, función de autocorrelación, función de autocovarianza, procesos estacionarios, procesos gaussianos, proceso de Poisson, sistemas lineales, densidad espectral de potencia.

11.1. Introducción

En muchos experimentos de tipo aleatorio el resultado es una función del tiempo (o del espacio).

Por ejemplo,

- en sistemas de reconocimiento de voz las decisiones se toman sobre la base de una onda que reproduce las características de la voz del interlocutor, pero la forma en que el mismo interlocutor dice una misma palabra sufre ligeras variaciones cada vez que lo hace;
- en un sistema de cola, por ejemplo, en un servidor de telecomunicaciones, el número de clientes en el sistema a la espera de ser atendidos evoluciona con el tiempo y está sujeto a condiciones tales que su comportamiento es *impredecible*;
- en un sistema de comunicación típico, la señal de entrada es una onda que evoluciona con el tiempo y que se introduce en un canal donde es contaminada por un ruido aleatorio, de tal manera que es imposible separar cuál es el mensaje original con absoluta *certeza*.
- ...

Desde un punto de vista matemático, todos estos ejemplos tienen en común que el fenómeno puede ser visto como unas funciones que dependen del tiempo, pero que son desconocidas a priori, porque dependen del *azar*. En este contexto vamos a definir el concepto de proceso aleatorio. Nuestro objetivo, como en capítulos anteriores dedicados a variables y vectores aleatorios, es describir desde un punto de vista estadístico el fenómeno, proporcionando medidas de posición, medidas sobre la variabilidad, etc.

11.1.1. Definición

Consideremos un experimento aleatorio sobre un espacio muestral Ω. Supongamos que para cada resultado posible, A, tenemos una observación del fenómeno dada por una función real de variable real, $x(t, A)$, con $t \in I \subset \mathbf{R}$. Habitualmente, t representa al tiempo, pero también puede referirse a otras magnitudes físicas. Para cada A vamos a denominar a $x(t, A)$ **realización** o **función muestral**.

Obsérvese que para cada $t_0 \in I$, $X(t, \cdot)$ es una variable aleatoria. Pues bien, al conjunto

$$\{X(t, A) : t \in I, A \in \Omega\}$$

lo denominamos **proceso aleatorio (en adelante p.a.) o estocástico**.

Si recordamos las definiciones de variable aleatoria y vector aleatorio, podemos ver en qué sentido están relacionados los conceptos de variable, vector y proceso aleatorio. Concretamente, si Ω es un espacio muestral, una variable aleatoria es una función

$$X : \Omega \to \mathbf{R}$$

que a cada suceso posible le asigna **un número real**. Por su parte, un vector aleatorio es básicamente una función

$$X : \Omega \to \mathbf{R}^N$$

que a cada suceso posible le asigna **un vector real**. Finalmente, un proceso aleatorio es básicamente una función

$$X : \Omega \to \{\text{funciones reales de vble real}\}$$

que a cada suceso posible le asigna **una función real**.

De cara a escribir de ahora en adelante un p.a., lo notaremos normalmente, por ejemplo, como $X(t)$, obviando así la variable que hace referencia al elemento del espacio muestral al que va asociada la función muestral. Este convenio es el mismo que nos lleva a escribir X refiriéndonos a una v.a. o a un vector.

11.1.2. Tipos de procesos aleatorios

El tiempo es una magnitud física intrínsecamente continua, es decir, que puede tomar cualquier valor de los números reales. Sin embargo, no siempre es posible observar las cosas *en cada instante del tiempo*. Por eso, en el ámbito de los procesos (no sólo estocásticos) es importante preguntarse si el fenómeno que representa el proceso es observado *en cada instante* o sólo *en momentos concretos del tiempo*.

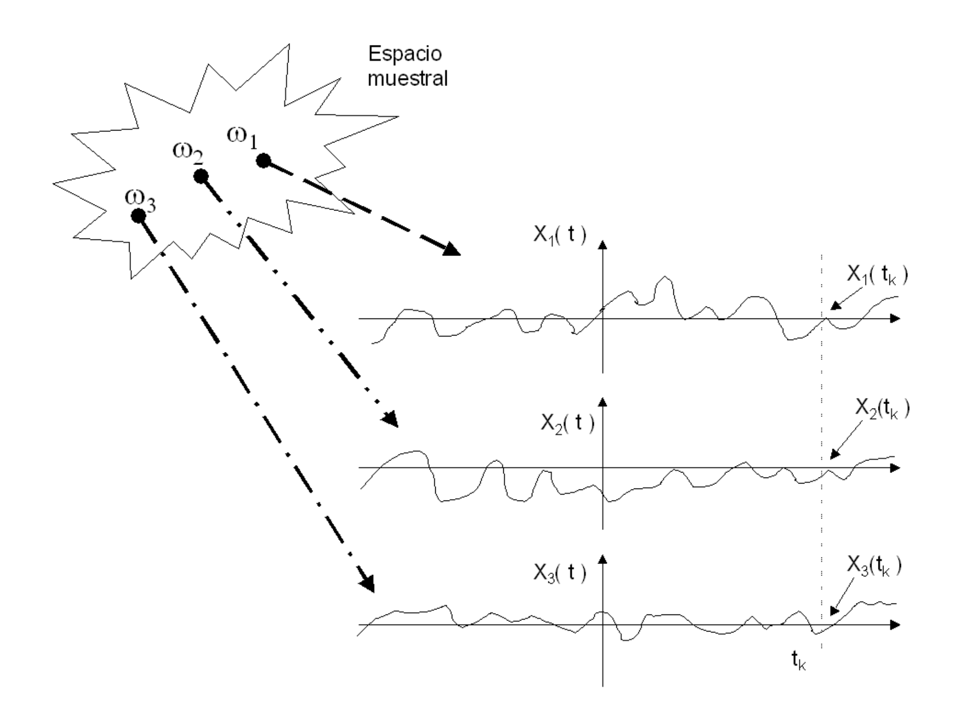


Figura 11.1: Representación de un proceso aleatorio.

Dado un espacio muestral Ω y un p.a. definido en él,

$$\{X(t, A) : t \in I, A \in \Omega\},$$

se dice que el proceso es un **p.a. en tiempo discreto** si I es un conjunto numerable.

En el caso de procesos en tiempo discreto se suele escribir X_n o $X[n]$ refiriéndonos a la notación más general $X(n)$. Por otra parte, el conjunto I normalmente es el conjunto de los enteros o de los enteros positivos, aunque también puede ser un subconjunto de éstos.

En algunos libros los procesos en tiempo discreto también son denominados **secuencias aleatorias**.

Dado un espacio muestral Ω y un p.a. definido en él,

$$\{X(t, A) : t \in I, A \in \Omega\},$$

se dice que el proceso es un **p.a. en tiempo continuo** si I es un intervalo.

En el caso de procesos en tiempo continuo, I es normalmente el conjunto de los reales positivos o un subconjunto de éstos.

Si nos damos cuenta, esta primera clasificación de los p.a. la hemos hecho en función del carácter discreto o continuo del tiempo, es decir, del conjunto I. Existe otra clasificación posible en función de cómo son las variables aleatorias del proceso, discretas o continuas. Sin embargo, ambos tipos de procesos, con variables discretas o con variables continuas, pueden estudiarse casi siempre de forma conjunta. Por ello sólo distin-

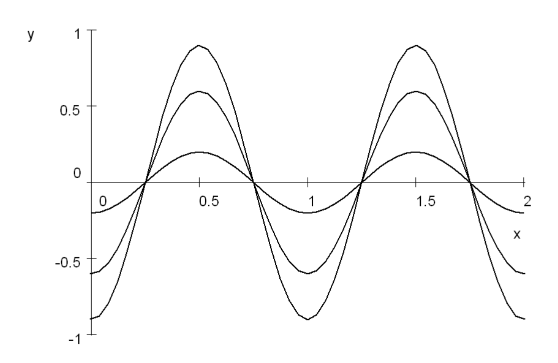


Figura 11.2: Distintas funciones muestrales de un proceso aleatorio.

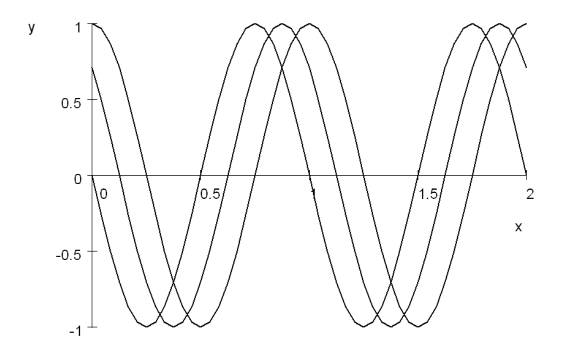


Figura 11.3: Distintas funciones muestrales de un proceso.

guiremos p.a. con variables discretas y p.a. con variables continuas si es necesario. En este sentido, cuando nos refiramos a la función masa (si el p.a. es de variables discretas) o a la función de densidad (si el p.a. es de variables continuas), hablaremos en general de función de densidad.

Ejemplo. Sea ξ una variable aleatoria uniforme en $(-1, 1)$. Definimos el proceso en tiempo continuo $X(t, \xi)$ como

$$X(t, \xi) = \xi \cos(2\pi t).$$

Sus funciones muestrales son ondas sinusoidales de amplitud aleatoria en $(-1, 1)$ (Figura 11.2).

Ejemplo. Sea θ una variable aleatoria uniforme en $(-\pi, \pi)$. Definimos el proceso en tiempo continuo $X(t, \pi)$ como

$$X(t, \pi) = \cos(2\pi t + \theta).$$

Sus funciones muestrales son versiones desplazadas aleatoriamente de $\cos(2\pi t)$ (Figura 11.3).

11.2. Descripción de un proceso aleatorio

11.2.1. Descripción estadística mediante distribuciones multidimensionales

En general, para especificar cómo es un p.a. de forma precisa es necesario caracterizar la distribución de probabilidad de cualquier subconjunto de variables del proceso. Es decir, si $X(t)$ es un p.a., es necesario conocer cuál es la distribución de cualquier vector del tipo

$$(X(t_1), ..., X(t_k)),$$

para todo $k > 0$, $(t_1, ..., t_k) \subset I$, mediante su función de distribución conjunta

$$F_{X(t_1),...,X(t_k)}(x_1, ..., x_k)$$

o mediante su función de densidad (o masa) conjunta

$$f_{X(t_1),...,X(t_k)}(x_1, ..., x_k).$$

Sin embargo, no siempre es fácil conocer todas las posibles distribuciones de todos los posibles vectores de variables del proceso. Por ello, para tener una descripción más sencilla aunque puede que incompleta del proceso, se acude a las medias, a las varianzas y a las covarianzas de sus variables.

11.2.2. Función media y funciones de autocorrelación y autocovarianza

Sea un p.a. $X(t)$. Se define la **función media** o simplemente la media de $X(t)$ como

$$\bar{X}(t) = \bar{x}(t) = E[X(t)] = \int_{-\infty}^{\infty} x f_{X(t)}(x)\, dx,$$

para cada $t \in I$.

Nótese que, como su nombre indica, se trata de una función determinística. No tiene ninguna componente aleatoria. Nótese también que aunque se está escribiendo el símbolo integral, podríamos estar refiriéndonos a una variable discreta, en cuyo caso se trataría de una suma.

Se define la **función de autocovarianza** o simplemente la **autocovarianza** de $X(t)$ como

$$\begin{aligned} C_X(t,s) &= Cov[X(t), X(s)] = E[(X(t) - m_X(t))\,(X(s) - m_X(s))] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_1 - \bar{x}(t))(x_2 - \bar{x}(s))\, f_{X(t),X(s)}(x_1, x_2)\, dx_2 dx_1 \end{aligned}$$

Se define la **función de autocorrelación** o simplemente la **autocorrelación** de $X(t)$ como

$$R_X(t,s) = E[X(t)\cdot X(s)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f_{X(t),X(s)}(x_1,x_2)\,dx_2 dx_1$$

Nótese, de cara al cálculo, que la diferencia entre ambas funciones tan sólo es el producto de las medias[1].

$$C_X(t,s) = R_X(t,s) - m_X(t)\cdot m_X(s).$$

De hecho, si el proceso está **centrado en media**, es decir, si su media es constantemente cero, ambas funciones coinciden.

Por otra parte, la varianza de las variables del proceso puede obtenerse como

$$Var(X(t)) = C_X(t,t).$$

La interpretación de la función de autocovarianza $C_X(t,s)$ es la de una función que proporciona una medida de la interdependencia lineal entre dos v.a. del proceso, $X(t)$ y $X(s)$, que distan $\tau = s - t$ unidades de tiempo. De hecho, ya sabemos que podríamos analizar esta relación mediante el coeficiente de correlación lineal

$$\rho_X(t,s) = \frac{C_X(t,s)}{\sqrt{C_X(t,t)\,C_X(s,s)}}.$$

Aparentemente es esperable que tanto más rápidamente cambie el proceso, más decrezca la autocorrelación conforme aumenta τ, aunque por ejemplo, los procesos periódicos no cumplen esa propiedad.

En el campo de la teoría de la señal aletatoria, a partir de la función de autocorrelación se puede distinguir una señal cuyos valores cambian muy rápidamente frente a una señal con variaciones más suaves. En el primer caso, la función de autocorrelación y de autocovarianza en instantes t y $t+\tau$ decrecerán lentamente con τ, mientras que en el segundo, ese descenso será mucho más rápido. En otras palabras, cuando la autocorrelación (o la autocovarianza) es alta, entre dos instantes cercanos del proceso tendremos valorer similares, pero cuando es baja, podremos tener fuertes diferencias entre valores cercanos en el tiempo.

La gran importancia de estas funciones asociadas a un proceso, media y autocovarianza (o autocorrelación), es por tanto que aportan toda la información acerca de la relación lineal que existe entre dos v.a. cualesquiera del proceso. Como hemos dicho, en la práctica, resulta extremadamente complicado conocer completamente la distribución de un proceso y, cuando esto ocurre, no siempre es sencillo utilizar las técnicas del cálculo de probabilidades para el tratamiento de estos procesos. Sin embargo, tan sólo con la información dada por la función media y la función de autocorrelación pueden ofrecerse resultados muy relevantes acerca de los procesos, tal y como hemos visto en el caso de variables y vectores aleatorios.

Ejemplo. La señal recibida por un receptor AM de radio es una señal sinusoidal con fase aleatoria, dada por $X(t) = A\cdot\cos(2\pi f_c t + \Xi)$, donde A y f_c son constantes y Ξ es una v.a. uniforme en $(-\pi,\pi)$.

[1] Esta fórmula es la misma que cuando veíamos la covarianza entre dos variables, calculable como *la media del producto menos el producto de las medias.*

En ese caso,

$$E[X(t)] = \int_{-\pi}^{\pi} A\cos(2\pi f_c t + \xi)\frac{1}{2\pi}d\xi = \frac{A}{2\pi}\left[\sin(2\pi f_c t + \xi)\right]_{\xi=-\pi}^{\xi=\pi}$$

$$= \frac{A}{2\pi}\left(\sin(2\pi f_c t)\cos(\pi) + \cos(2\pi f_c t)\sin(\pi) - \sin(2\pi f_c t)\cos(-\pi) - \cos(2\pi f_c t)\sin(-\pi)\right)$$
$$= \frac{A}{2\pi}[0+0] = 0.$$

$$R_X(t,t+\tau) = E[X(t+\tau)X(t)] = E\left[A^2\cos(2\pi f_c t + 2\pi f_c \tau + \Xi)\cos(2\pi f_c t + \Xi)\right]$$
$$= \frac{A^2}{2}E[\cos(4\pi f_c t + 2\pi f_c \tau + 2\Xi)] + \frac{A^2}{2}E[\cos(2\pi f_c \tau)]$$

$$= \frac{A^2}{2}\int_{-\pi}^{\pi}\frac{1}{2\pi}\cos(4\pi f_c t + 2\pi f_c \tau + 2\xi)\,d\xi + \frac{A^2}{2}\cos(2\pi f_c \tau)$$
$$= \frac{A^2}{2}\cdot 0 + \frac{A^2}{2}\cos(2\pi f_c \tau) = \frac{A^2}{2}\cos(2\pi f_c \tau).$$

Por tanto,

$$C_X(t,t+\tau) = R_X(t,t+\tau) - m_X(t)\,m_X(t+\tau) = \frac{A^2}{2}\cos(2\pi f_c \tau).$$

11.3. Tipos más comunes de procesos aleatorios

En este apartado definimos propiedades que pueden ser verificadas por algunos procesos aleatorios y que les confieren características especiales en las aplicaciones prácticas.

11.3.1. Procesos independientes

Sea un p.a. $X(t)$. Si para cada n instantes de tiempo, $t_1,...,t_n$, las v.a. del proceso en esos instantes son independientes, es decir,

$$f_{X(t_1),...,X(t_n)}(x_1,...,x_n) = f_{X(t_1)}(x_1)\cdot ...\cdot f_{X(t_n)}(x_n),$$

se dice que el proceso es **independiente.**

La interpretación de este tipo de procesos es la de aquellos en donde el valor de la v.a. que es el proceso en un momento dado no tiene nada que ver con el valor del proceso en cualquier otro instante. Desde un punto de vista físico estos procesos son muy *caóticos* y se asocian en la práctica a ruidos que no guardan en un momento dado ninguna relación consigo mismos en momentos adyacentes.

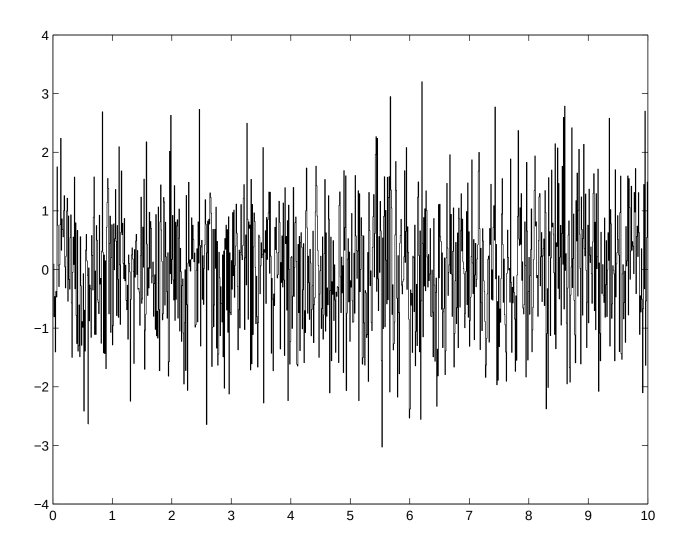


Figura 11.4: Función muestral de un proceso independiente formado por v.a gaussianas de media cero y varianza uno.

11.3.2. Procesos con incrementos independientes

Sea un p.a. $X(t)$. Se dice que tiene incrementos independientes si cualquier conjunto de N v.a. del proceso, $X(t_1), X(t_2), ..., X(t_N)$, con $t_1 < t_2 < ... < t_N$ son tales que los incrementos

$$X(t_1), X(t_2) - X(t_1), ..., X(t_N) - X(t_{N-1})$$

son independientes entre sí.

11.3.3. Procesos de Markov

No debemos perder de vista la complejidad que implica la descripción estadística de un proceso aleatorio. Pensemos por ejemplo que un proceso ha evolucionado hasta un instante t y se conoce esa evolución; es decir, se conoce el valor $X(s) = x_s$ para todo $s \leq t$. Si se desea describir la posición del proceso en un instante posterior a t, $t + \Delta$, sería necesario calcular la distribución condicionada

$$X(t+\Delta) \mid \{X(s) = x_s \text{ para todo } s \leq t\}.$$

Esto, en general, es bastante complejo.

Además, ¿tiene sentido pensar que la evolución del proceso en el instante $t + \Delta$ se vea afectada por toda la historia del proceso, desde el instante inicial $s = 0$ hasta el último instante de esa historia $s = t$? Parece lógico pensar que la evolución del proceso tenga en cuenta la historia más reciente de éste, pero no toda la historia. Esta hipotesis se ve avalada por los perfiles más habituales de las funciones de autocorrelación, donde observamos que la relación entre variables del proceso suele decrecer en la mayoría de las ocasiones conforme aumenta la distancia en el tiempo entre las mismas.

Los procesos de Markov son un caso donde esto ocurre. Se trata de procesos que evolucionan de manera que en cada instante *olvidan* todo su pasado y sólo tienen en cuenta para su evolución futura el instante más

reciente, más actual. En el siguiente sentido:

Un proceso $X(t)$ se dice **markoviano o de Markov** si para cualesquiera $t_1 < ... < t_n < t_{n+1}$ instantes consecutivos de tiempo se verifica

$$f_{X(t_{n+1})|X(t_1)=x_1,...,X(t_n)=x_n}(x_{n+1}) = f_{X(t_{n+1})|X(t_n)=x_n}(x_{n+1}).$$

Esta definición se suele enunciar coloquialmente diciendo que un proceso de Markov es *aquel cuyo futuro no depende del pasado sino tan sólo del presente.*

11.3.4. Procesos débilmente estacionarios

Una de las propiedades más usuales en los procesos estocásticos consiste en una cierta estabilidad en sus medias y en sus covaranzas, en el sentido en que vamos a describir a continuación.

$X(t)$ es un proceso **débilmente estacionario** si
$m_X(t)$ es independiente de t y
$C(t,s)$ (o $R(t,s)$) depende tan sólo de $s-t$, en cuyo caso se nota $C(s-t)$ (ó $R(s-t)$).

Es importante destacar que la primera de las condiciones es irrelevante, ya que siempre se puede centrar en media un proceso para que ésta sea cero, constante. Es decir, en la práctica es indiferente estudiar un proceso $X(t)$ con función media $\mu_X(t)$ que estudiar el proceso $Y(t) = X(t) - \mu_X(t)$, con media cero.

La propiedad más exigente y realmente importante es la segunda. Viene a decir que la relación entre variables aleatorias del proceso sólo depende de la distancia en el tiempo que las separa.

Nota. Vamos a hacer una puntualización muy importante respecto a la notación que emplearemos en adelante. Acabamos de ver que si un proceso es débilmente estacionario, sus funciones de autocovarianza y de autocorrelación, $C(s,t)$ y $R(s,t)$ no dependen en realidad de s y de t, sino tan sólo de $t-s$. Por eso introducimos la notación

$$C(t,s) \equiv C(s-t)$$
$$R(t,s) = R(s-t).$$

Por lo tanto, ¿qué queremos decir si escribimos directamente $C(\tau)$ o $R(\tau)$? Que tenemos un p.a. débilmente estacionario y que hablamos de

$$C(\tau) = C(t,t+\tau)$$
$$R(\tau) = R(t,t+\tau).$$

Una medida importante asociada a un proceso débilmente estacionario es la **potencia promedio**, definida como la media del cuadrado de éste en cada instante t, es decir $R_X(0) = E\left[|X(t)|^2\right]$. Más adelante observaremos con detenimiento esta medida.

Por otra parte, la peculiaridad que define a los procesos débilmente estacionarios le confiere a su función de autocorrelación y autocovarianza dos propiedades interesantes: sea $X(t)$ un proceso estacionario (débil). Entonces, si notamos $R_X(\tau) = E[X(t)X(t+\tau)]$ para todo t, su función de autocorrelación y por $C_X(\tau)$ a su función de autocovarianza:

1. Ambas son funciones pares, es decir, $R_X(-\tau) = R_X(\tau)$ y $C_X(-\tau) = C_X(\tau)$.

2. $|R_X(\tau)| \leq R_X(0)$ y $|C_X(\tau)| \leq C_X(0) = \sigma^2$ para todo τ.

Ejemplo. En el ejemplo del oscilador vimos que la señal recibida por un receptor AM de radio es una señal sinusoidal con fase aleatoria, dada por $X(t) = A \cdot \cos(2\pi f_c t + \Xi)$, donde A y f_c son constantes y Ξ es una v.a. uniforme en $(-\pi, \pi)$ tiene por función media

$$E[X(t)] = 0$$

y por función de autocorrelación

$$R_X(t, t+\tau) = \frac{A^2}{2}\cos(2\pi f_c \tau).$$

De esta forma, podemos ver que el proceso es débilmente estacionario.

Ejemplo. Un proceso binomial es un proceso con función de autocovarianza

$$C(m,n) = \text{mín}(m,n)\, p(1-p),$$

que no depende sólo de $m-n$. Por lo tanto no es débilmente estacionario.

Ejemplo. Vamos a considerar un proceso en tiempo discreto e independiente, X_n, con media cero y varianza constante e igual a σ^2. Vamos a considerar también otro proceso que en cada instante de tiempo considera la media de X en ese instante y el anterior, es decir,

$$Y_n = \frac{X_n + X_{n-1}}{2}.$$

En primer lugar, dado que $E[X_n] = 0$ para todo n, lo mismo ocurre con Y_n, es decir,

$$E[Y_n] = E\left[\frac{X_n + X_{n-1}}{2}\right] = 0.$$

Por otra parte,

$$\begin{aligned} C_Y(n,n+m) &= R_Y(n,n+m) - 0 = E[Y(n)Y(n+m)] \\ &= E\left[\frac{X_n + X_{n-1}}{2}\frac{X_{n+m} + X_{n+m-1}}{2}\right] \\ &= \frac{1}{4}E[(X_n + X_{n-1})(X_{n+m} + X_{n+m-1})] \\ &= \frac{1}{4}(E[X_nX_{n+m}] + E[X_nX_{n+m-1}] + E[X_{n-1}X_{n+m}] + E[X_{n-1}X_{n+m-1}]) \end{aligned}$$

Ahora debemos tener en cuenta que

$$C_X(n,m) = R_X(n,m) = \begin{cases} 0 & \text{si } n \neq m \\ \sigma^2 & \text{si } n = m \end{cases},$$

ya que X_n es un proceso independiente. Por lo tanto,

$$\begin{aligned} C_Y(n,n+m) &= \begin{cases} \frac{1}{4}\left(\sigma^2 + 0 + 0 + \sigma^2\right) & \text{si } m = 0 \\ \frac{1}{4}\left(0 + \sigma^2 + 0 + 0\right) & \text{si } m = 1 \\ \frac{1}{4}\left(0 + 0 + \sigma^2 + 0\right) & \text{si } m = -1 \\ 0 & \text{en otro caso} \end{cases} \\ &= \begin{cases} \frac{1}{2}\sigma^2 & \text{si } m = 0 \\ \frac{1}{4}\sigma^2 & \text{si } m = \pm 1 \\ 0 & \text{en otro caso} \end{cases} \end{aligned}$$

Podemos decir, por tanto, que el proceso Y_n también es débilmente estacionario, porque su media es constante (cero) y $C_Y(n,n+m)$ no depende de n sino tan sólo de m.

11.3.5. Procesos ergódicos

Si nos damos cuenta, estamos describiendo los procesos aleatorios a partir de promedios estadísticos, principalmente a partir de la media de cada una de sus variables y de sus correlaciones. Vamos a centrarnos en procesos débilmente estacionarios. En ese caso, los promedios estadísticos más relevantes serían la media,

$$E[X(t)] = m_X(t) = m_X = \int_{-\infty}^{\infty} x f_{X(t)}(x)\,dx$$

y la autocorrelación entre dos variables que disten τ unidades de tiempo,

$$R_X(\tau) = E[X(t)X(t+\tau)] = \int_{-\infty}^{\infty} x_1x_2 f_{X(t)X(t+\tau)}(x_1,x_2)\,dx_1dx_2.$$

Hasta ahora quizá no lo habíamos pensado, pero más allá de los típicos ejemplos, ¿cómo podríamos tratar de calcular o estimar al menos estas cantidades? Si aplicamos lo que hemos aprendido hasta ahora, estimaríamos, por ejemplo, la media con la media muestral, pero para ello necesitaríamos una muestra muy grande de

funciones muestrales del proceso, y eso no siempre ocurre. De hecho, no es nada rara la situación en la que, en realidad, sólo es posible observar una única función muestral del proceso.

Ahora bien, dada una única función muestral de un proceso, $x(t)$, en esa función hay muchos datos, tantos como instantes de tiempo t hayamos sido capaces de observar. ¿No podría ocurrir que utilizáramos todos esos datos que hay en $x(t)$para estimar las medias y las autocorrelaciones? Por ejemplo, si tenemos observada la señal $x(t)$ en un montón de valores $t_1, ...t_n$, ¿qué tendrá que ver

$$\frac{x(t_1) + ...x(t_n)}{n}$$

con la media del proceso m_X? De hecho, si n es muy grande y corresponde a un intervalo de observación $[-T, T]$, tendríamos que

$$\frac{x(t_1) + ... + x(t_n)}{n} \simeq \frac{1}{2T} \int_{-T}^{T} x(t)\, dt.$$

Ahora no es una integral sobre los valores de x (integral *estadística*) sino sobre el tiempo.

En el caso de la autocorrelación pasaría igual, tendríamos que podríamos observar un montón de pares de valores de la señal en los instantes $t_1, ..., t_n$ y $t_1 + \tau, ..., t_n + \tau$ en el intervalo $[-T, T]$ y con ellos podríamos estimar

$$\frac{1}{2T} \int_{-T}^{T} x(t)\, x(t+\tau)\, dt \simeq \frac{x(t_1)\, x(t_1+\tau) + ... + x(t_n)\, x(t_n+\tau)}{n}.$$

Lo que no sabemos, en general, es si esa integral tiene algo que ver con $R_X(\tau)$, que es una integral *estadística*.

Pues bien, se dice que un proceso estacionario es **ergódico** cuando las funciones que entrañan valores esperados a lo largo de las realizaciones (integrales o promedios *estadísticos*) pueden obtenerse también a partir de una sola función muestral $x(t)$. Es decir, que una sola realización es representativa de todo el proceso. Más concretamente, un proceso será ergódico en media y en autocorrelación si

$$lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)\, dt = m_X$$

y

$$lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x(t)\, x(t+\tau)\, dt = R_X(\tau).$$

11.4. Ejemplos de procesos aleatorios

11.4.1. Ruidos blancos

En telecomunicaciones los ruidos son señales que se adhieren a la señal enviada en cualquier proceso de comunicación, de tal manera que uno de los objetivos fundamentales en este tipo de procesos es, dada la señal resultante de sumar la señal enviada, $X(t)$, y el ruido del canal, $N(t)$, es decir, dada $Y(t) = X(t) + N(t)$, saber *filtrar* esta señal para estimar cuál es el verdadero valor de $X(t)$.

En este apartado nos referimos brevemente a un modelo gastante común para los fenómenos de ruido, llamado ruido blanco.

Un **ruido blanco** es un proceso $N(t)$ centrado, débilmente estacionario e incorrelado con varianza $\frac{N_0}{2}$. Por tanto, su función de autocovarianza (y autocorrelación) será

$$C_N(t, t+\tau) = \begin{cases} \frac{N_0}{2} & \text{si } \tau = 0 \\ 0 & \text{en otro caso} \end{cases}.$$

Utilizando la llamada función impulso, dada por

$$\delta(t) = \begin{cases} 1 & \text{si } t = 0 \\ 0 & \text{en otro caso} \end{cases},$$

esta función de autocovarianza puede escribirse como

$$C_N(\tau) = \frac{N_0}{2}\delta(\tau).$$

La justificación de que este sea un modelo habitual para los ruidos, considerando que los valores del ruido están incorrelados unos con otros, es que suelen ser debidos a fenómenos completamente aleatorios y caóticos, por lo que no es esperable que exista relación entre valores del ruido, ni siquiera cuando éstos son muy cercanos en el tiempo.

11.4.2. Procesos gaussianos

Hasta ahora hemos definido y estudiado familias muy genéricas de procesos (independientes, estacionarios, ...). En esta sección vamos a considerar más concretamente la conocida como familia de procesos aleatorios gaussianos, que constituye, sin duda, la más importante de entre las que se utilizan en Telecomunicaciones y en cualquier otro ámbito de aplicación de la Estadística.

Un p.a. $X(t)$ se dice **proceso gaussiano** si cualquier colección de variables del proceso tiene distribución conjuntamente gaussiana. Es decir, si cualquier colección $X(t_1), ..., X(t_n)$ tiene función de densidad conjunta

$$f_{X(t_1),...,X(t_n)}(x_1, ..., x_n) = \frac{1}{\sqrt{(2\pi)^n \det(C)}} \exp\left[-\frac{1}{2}(x-\mu)' \cdot C^{-1} \cdot (x-\mu)\right],$$

donde

$$\begin{aligned} x &= (x_1, ..., x_n)', \\ \mu &= (E[X(t_1)], ..., E[X(t_n)])', \\ C &= (C_{i,j})_{i,j=1,..,n}, \\ C_{ij} &= Cov[X(t_i), X(t_j)]. \end{aligned}$$

Nótese que un proceso gaussiano está completamente descrito una vez que se conocen su función media y su autocovarianza o su autocorrelación.

Existen dos razones fundamentales por las que, como hemos comentado, los procesos gaussianos son la familia de procesos más relevante:

- Por una parte, las propiedades analíticas que verifican los hacen fácilmente manejables, como veremos a continuación.
- Por otra parte, estos procesos han demostrado ser un excelente modelo matemático para gran número de experimentos o fenómenos reales (resultado amparado en el Teorema Central del Límite).

Ejemplo. Es muy habitual considerar que los ruidos blancos son gaussianos. En ese caso, si consideramos ruidos blancos gaussianos, sus variables no sólo son incorreladas, sino que también son independientes.

Ejemplo. Sea un proceso gaussiano $X(t)$ débilmente estacionario con $E[X(t)] = 4$ y autocorrelación $R_X(\tau) = 25e^{-3|\tau|} + 16$. Obsérvese que la autocorrelación (y la autocovarianza) decrece rápidamente con el paso del tiempo.

Si deseamos caracterizar la distribución de probabilidad de tres v.a. del proceso, observadas en los instantes $t_0, t_1 = t_0 + \frac{1}{2}$ y $t_2 = t_1 + \frac{1}{2} = t_0 + 1$, necesitamos las medias, $E[X(t_i)] = 4$ y la matriz de covarianzas, dada a partir de $C_X(\tau) = 25e^{-3|\tau|}$.

$$C_{X(t_0),X(t_1),X(t_2)} = \begin{pmatrix} 25 & 25e^{-3/2} & 25e^{-6/2} \\ 25e^{-3/2} & 25 & 25e^{-3/2} \\ 25e^{-6/2} & 25e^{-3/2} & 25 \end{pmatrix}.$$

Algunas propiedades de interés de los procesos gaussianos:

- Un proceso gaussiano es independiente si y sólo si $C(t_i, t_j) = 0$ para todo $i \neq j$.
- Sea $X(t)$ un proceso gaussiano. Este proceso es markoviano si y sólo si

  $$C_X(t_1, t_3) = \frac{C_X(t_1, t_2) \cdot C_X(t_2, t_3)}{C_X(t_2, t_2)},$$

  para cualesquiera $t_1 < t_2 < t_3$.
- Un proceso $X(t)$ gaussiano, centrado, con incrementos independientes y estacionarios es de Markov.

11.4.3. Procesos de Poisson

El proceso de Poisson es un modelo para procesos de la vida real que cuentan ocurrencias de un suceso a lo largo del tiempo, denominados por ello *procesos de recuento*.

Algunos de los ejemplos más comunes en el campo de las Telecomunicaciones son el proceso que cuenta el número de llamadas recibidas en una centralita telefónica o el que cuenta el número de visitas a una página WEB. En otros ámbitos, como la Física, estos procesos pueden servir, por ejemplo, para contabilizar el número de partículas emitidas por un cuerpo.

En todas estas aplicaciones, el proceso tendría la expresión

$$N(t) = \sum_{n=1}^{\infty} u(t - T[n]),$$

donde $T[n]$ es un proceso en tiempo discreto que representa el momento de la $n-$ésima llegada que cuenta el proceso y

$$u(t - t_0) = \begin{cases} 0 & \text{si } t < t_0 \\ 1 & \text{si } t \geq t_0 \end{cases}$$

es la función umbral.

El **proceso de Poisson de parámetro** λ es el proceso $N(t) = \sum_{n=1}^{\infty} u(t - T[n])$ para el cual la v.a. $T[n]$ es una suma de n exponenciales independientes del mismo parámetro λ, lo que genera una distribución de Erlang de parámetros n y λ, con función de densidad

$$f_{T[n]}(t) = \frac{(\lambda t)^{n-1}}{(n-1)!} \lambda e^{-\lambda t} u(t).$$

Alternativamente, puede decirse que **el proceso de Poisson es aquél en el que los tiempos entre llegadas,**

$$\Upsilon[n] = T[n] - T[n-1],$$

siguen siempre distribuciones exponenciales independientes[a] **del mismo parámetro,** esto es

$$f_{\Upsilon[n]}(t) = \lambda e^{-\lambda t} u(t).$$

[a] Obsérvese por tanto que el proceso $T[n]$ tiene incrementos independientes.

Ejemplo. En la Figura 11.6 se muestran funciones muestrales de un proceso de Poisson de parámetro $\lambda = 1$. Vamos a interpretar la función muestral de la izquierda pensando, por ejemplo, que representa el número de visitas a una página WEB: se observa que poco depués de los tres minutos se han dado 3 visitas; después pasan casi 5 minutos sin ninguna visita; a continuación se producen un buen número de visitas en poco tiempo; ...

Si observamos tan sólo el eje del tiempo, podríamos señalar los instantes en que se producen las llegadas. Sabemos que esos incrementos en el tiempo desde que se produce una llegada hasta la siguiente siguen una distribución exponencial, en este caso de parámetro 1.

Vamos a describir algunas de las propiedades más interesantes de los procesos de Poisson:

- Sea $N(t)$ un proceso de Poisson de parámetro λ. Entonces, para todo t se tiene que $N(t) \to P(\lambda t)$.
- La media de un proceso de Poisson de parámetro λ es $\mu_N(t) = \lambda t$. Por tanto, el proceso de Poisson no es estacionario.
- Sea $N(t)$ un proceso de Poisson de parámetro λ. Entonces, el proceso tiene incrementos independientes

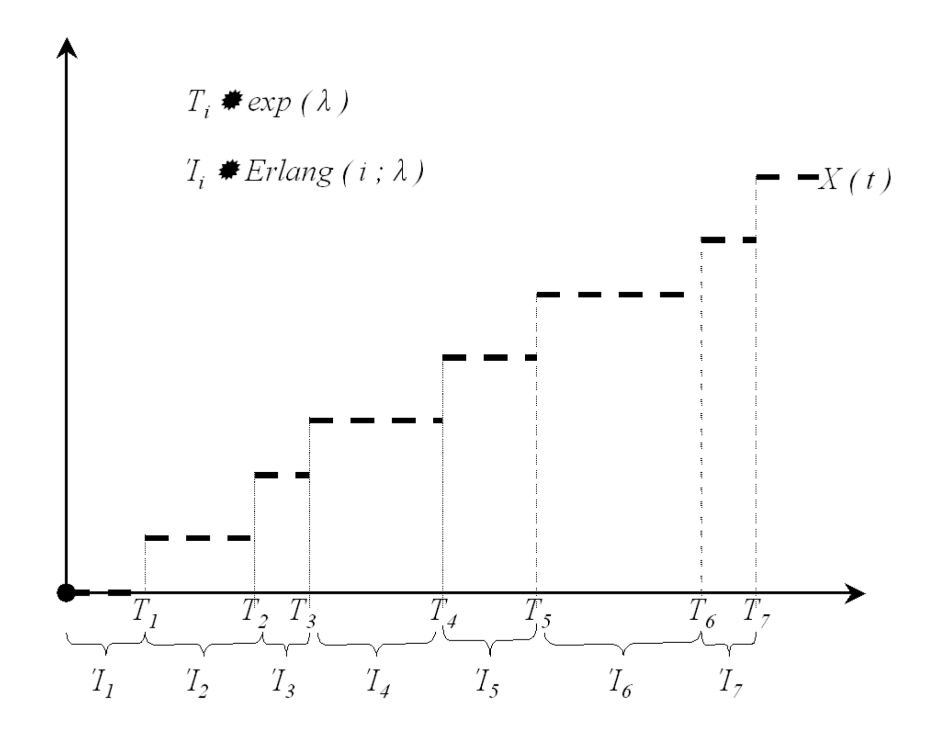


Figura 11.5: Representación gráfica de una función muestral de un p.a. de Poisson.

y para cualesquiera $t_1 < t_2$, el incremento $N(t_2) - N(t_1)$ sigue una distribución de Poisson de parámetro $\lambda(t_2 - t_1)$.

- Sea $N(t)$ un proceso de Poisson de parámetro λ. Entonces

$$C_N(t_1, t_2) = \lambda \operatorname{mín}(t_1, t_2).$$

- Sea $N(t)$ un proceso de Poisson de parámetro λ. Entonces, para cualesquiera $t_1 < ... < t_k$,

$$f_{N(t_1),...,N(t_k)}(n_1, ..., n_k)$$
$$= \begin{cases} e^{-\alpha_1} \frac{\alpha_1^{n_1}}{n_1!} \cdot e^{-\alpha_2} \frac{\alpha_2^{n_2-n_1}}{(n_2-n_1)!} \cdot ... \cdot e^{-\alpha_k} \frac{\alpha_2^{n_k-n_{k-1}}}{(n_k-n_{k-1})!} \text{ si } n_1 \leq ... \leq n_k \\ 0 \text{ en otro caso} \end{cases},$$

donde $\alpha_i = \lambda(t_i - t_{i-1})$.

- El proceso de Poisson es de Markov.

- Sean $N_1(t)$ p.a. de Poisson de parámetro λ_1, $N_2(t)$ p.a. de Poisson de parámetro λ_2, ambos independientes. Entonces, $N_1(t) + N_2(t)$ es un p.a. de Poisson de parámetro $\lambda_1 + \lambda_2$. Esta propiedad se conoce como *propiedad aditiva.*

- Sea $N(t)$ un p.a. de Poisson de parámetro λ. Supongamos que de todos los eventos que cuenta el proceso, sólo consideramos una parte de ellos; concretamente los que presentan una característica que tiene probabilidad p entre todos los eventos. En ese caso, si notamos por $N_p(t)$ al proceso que cuenta

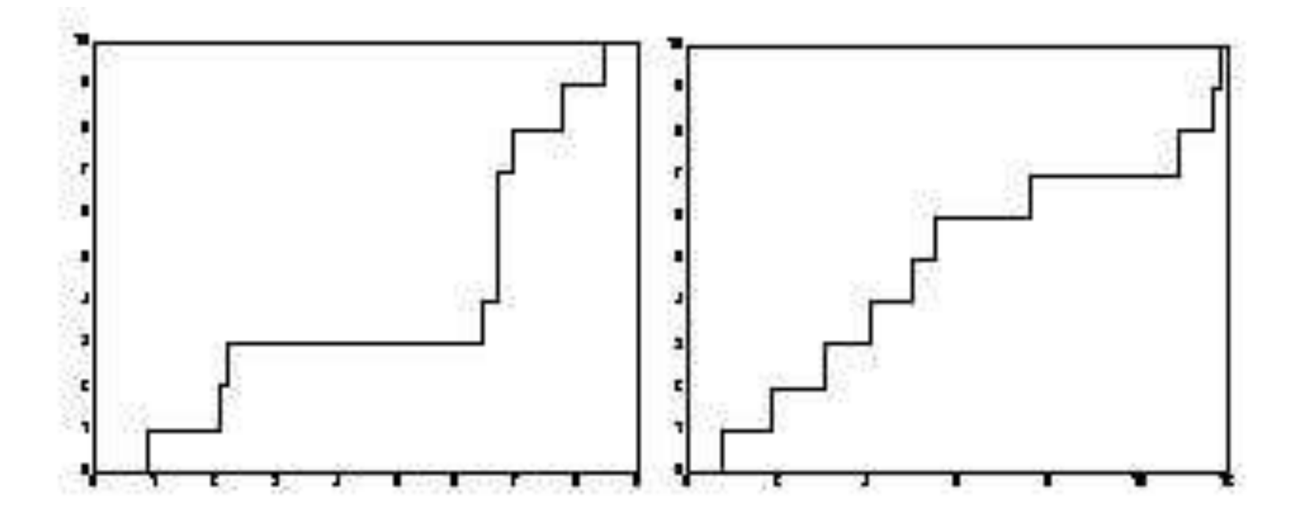


Figura 11.6: Funciones muestrales de un proceso de Poisson de parámetro 1.

los eventos con la característica dada, dicho proceso es de Poisson de parámetro $\lambda \cdot p$. Esta propiedad se conoce como *propiedad de descomposición*.

- El tiempo W que transcurre desde un instante arbitrario t_0 hasta la siguiente discontinuidad de un proceso de Poisson de parámetro λ es una variable aleatoria exponencial de parámetro λ, independientemente de la elección del punto t_0. Esta propiedad aparentemente paradójica se conoce como **propiedad de no memoria** del proceso de Poisson. Obsérvese que, en realidad, esta propiedad de no memoria lo es de la distribución exponencial.

Ejemplo. Es frecuente considerar que el proceso que cuenta el número de partículas emitidas por un material radiactivo es un proceso de Poisson. Vamos a suponer por tanto, que estamos observando el comportamiento de un determinado material del que se conoce que emite a razón de λ partículas por segundo.

Supongamos que se observa el proceso que cuenta el número de partículas emitidas desde un instante t hasta el instante $t + T_0$. Si en ese intervalo de tiempo se supera un umbral de N_0 partículas, debería sonar una señal de alarma. En ese caso, la probabilidad de que la alarma suene es

$$P\left[N\left(t+T_0\right)-N\left(t\right)>N_0\right]=\sum_{k=N_0+1}^{\infty}e^{-\lambda T_0}\frac{\left(\lambda T_0\right)^k}{k!}=1-\sum_{k=0}^{N_0}e^{-\lambda T_0}\frac{\left(\lambda T_0\right)^k}{k!},$$

ya que $N\left(t+T_0\right)-N\left(t\right)\to P\left(\lambda T_0\right)$.

Ejemplo. El número de visitas a la página WEB de una empresa que desea vender sus productos a través de INTERNET es adecuadamente descrito mediante un proceso de Poisson. Sabiendo que durante una hora se reciben un promedio de 5 visitas,

1. ¿cuál es la probabilidad de que no se reciba ninguna visita en media hora?

$$P[N(0.5)=0]=e^{-5\times 0.5}\frac{(5\times 0.5)^0}{0!}=8.2085\times 10^{-2},$$

   apenas un 8 % de probabilidad.

2. ¿Cuál es el promedio de visitas en 5 horas a la WEB? $E[N(5)]=5\times 5=25\,visitas$.

3. La empresa absorbe otra empresa del sector y opta por establecer un enlace directamente desde la página de su filial a la propia, garantizándose que todos los clientes de la filial visitan su página. Si el promedio de clientes que visitaban la página de la filial era de 2 clientes a la hora, ¿cuál es la probabilidad de que tras la fusión no se reciba ninguna visita en 10 minutos?

   Al hacerse con los clientes de la otra empresa (notemos por $M(t)$ al proceso de Poisson que contaba sus visitas, de parámetro $\lambda=2$ $visitas/hora$), lo que ha ocurrido es que ahora el número de visitas a la WEB de la empresa es la suma de ambos procesos: $T(t)=N(t)+M(t)$.

   Suponiendo que los procesos de Poisson que contaban las visitas a ambas empresas fueran independientes, se tiene que $T(t)$, en virtud de la propiedad aditiva del proceso de Poisson, es también un proceso de Poisson, de parámetro $\lambda=5+2=7\,visitas/hora$. Por tanto,

$$P\left[T\left(\frac{1}{6}\right)=0\right]=e^{-7\times\frac{1}{6}}\frac{\left(7\times\frac{1}{6}\right)^0}{0!}=0.3114,$$

   una probabilidad del 31 %.

Bibliografía

[Canavos, G. C. (1988)] Canavos, G. C. (1988). Probabilidad y Estadística. Aplicaciones y Métodos. McGraw-Hill.

[DeVore, J. L. (2004)] DeVore, J. L. (2004). Probabilidad y estadística para ingeniería y ciencias (6ª edición). Thomson.

[Johnson, R. A. (1997)] Johnson, R. A. (1997). Probabilidad y estadística para Ingenieros (5ª edición). Prentice Hall.

[Leon-Garcia, A.] Leon-Garcia, A. (1994). Probability and Random Processes for Electrical Engineering (2nd edition). Addison-Wesley.

[Lipschutz, S. & Schiller, J. (2000)] Lipschutz, S. & Schiller, J. (2000). Introducción a la Probabilidad y la Estadística. McGraw-Hill.

[Mendenhal, W & Sincich, T. (1997)] Mendenhal, W & Sincich, T. (1997). Probabilidad y Estadística para Ingeniería y Ciencias (4ª edición). Prentice Hall.

[Montgomery, D. C. & Runger, G. C. (2002)] Montgomery, D. C. & Runger, G. C. (2002). Probabilidad y estadística aplicadas a la Ingeniería (2ª edición). Wiley.

[Navidi, W. (2006)] Navidi, W. (2006). Estadística para ingenieros y científicos. McGraw-Hill.

[Ross, S. M. (2005)] Ross, S. M. (2005). Introducción a la Estadística. Editorial Reverté.

[Spiegel et al. (2010)] Spiegel, M. R., Schiller, J. y Srinivasan, R. A. (2010). Probabilidad y estadística (3ª edición), serie Schaum. McGraw-Hill.

[Walpole, R. E *et al* (1998)] Walpole, R. E., Myers, R. H. & Myers, S. L. (1998). Probabilidad y Estadística para Ingenieros (6ª edición). Prentice Hall.

Índice alfabético